Universidade de Brasília

IE- Departamento de Estatística

Estágio Supervisionado II

Análise de desempenho acadêmico nas disciplinas Probabilidade e Estatística, Estatística Aplicada e Bioestatística.

Davidson Martins Pereira

Gabriela Veríssimo Teixeira

Relatório Final

Orientadora: Prof.ª Maria Teresa Leão Costa

Brasília

Dezembro de 2011

Epígrafe

“O segredo de qualquer conquista é a coisa mais simples do mundo: Saber o que fazer com ela.”

Autor: Desconhecido

Agradecimentos

Agradecemos primeiramente a Deus, depois aos nossos pais, pela dedicação e apoio incondicionais. Agradecemos também aos nossos familiares e amigos, pelo companheirismo e dedicação. Em especial, a professora Maria Teresa pela paciência, orientação e disponibilidade de tempo e vontade de nos ajudar sempre que necessário.

Resumo

Neste Presente Trabalho foi estudado o desempenho dos alunos da Universidade de Brasília (Campus Darcy Ribeiro) nas seguintes disciplinas ofertadas pelo Departamento de Estatística para os diferentes cursos como: Estatística aplicada, Bioestatísticas ou Probabilidade e Estatística. O intuito deste estudo é identificar características associadas à reprovação nestas disciplinas com base no banco de dados extraído do Sistema de Informação de Graduação.

Utilizando modelagem via regressão logística observou-se algumas características comuns na reprovação dos alunos em todas as disciplinas, como: “Sexo”, “Professor”, “Curso” “Período que cursou a disciplina” e “Período que ingressou na UnB”. Outras características foram particulares a cada disciplina, como “nacionalidade” a Estatística Aplicada, e se o aluno “fez a disciplina no fluxo” para Bioestatística e Probabilidade e Estatística.

Sumário

1 - Introdução e Justificativa

A maior parte dos cursos universitários tem pelo menos uma disciplina de Estatística. Na UnB, o Departamento de Estatística é responsável pela oferta destas disciplinas, atendendo, atualmente, 1370 estudantes em três disciplinas: Probabilidade e Estatística (553 matrículas), Bioestatística (173 matrículas) e Estatística Aplicada (644 matrículas) num total de 28 turmas com 50 estudantes em média.

Avaliar o rendimento dos discentes, verificar se existe diferença e que fatores estão associados a estas diferenças é importante para a melhoria do processo de ensino-aprendizagem destas disciplinas e, principalmente, para contribuir efetivamente na formação profissional destes estudantes.

Em uma universidade com grande número de alunos como a UnB, o acompanhamento do rendimento dos alunos também se faz necessário já que a demanda destas matérias comuns a vários cursos é alta e a reprovação nelas é notória.

Com este trabalho, deseja-se traçar um perfil dos estudantes do Campus Darcy Ribeiro que cursam disciplinas oferecidas pelo Departamento de Estatística para outros cursos (Probabilidade e Estatística, Estatística Aplicada e Bioestatística), a fim de se verificar se existe diferença de desempenho entre cursos, dentre outras variáveis. Para tal, o banco utilizado será extraído do Sistema de Informação de Graduação (SIGRA) abrangendo o primeiro semestre de 2004 ao segundo semestre de 2008.

Neste estudo será de suma importância o uso das técnicas de dados categorizados focando na regressão logística uma vez que a variável de interesse é binária (reprovado e aprovado).

Em uma variedade de problemas de regressão, a variável resposta de interesse tem duas possíveis saídas qualitativas, sendo esta na maioria das

vezes representadas por “0“ e “1”, ou seja, sucesso ou fracasso, associada a variáveis explicativas qualitativas e quantitativas.

O modelo regressivo ajustado a problemas em que a variável *“Y”* é binária fornece como resposta a probabilidade de ocorrência de um evento, que seria sua média $E(Y)$. Os parâmetros estimados deste possuem interpretações simples e tangíveis. Da mesma forma que os demais modelos, a Regressão Logística possibilita predições, estimativas, análise residual e uma possível validação do modelo.

No estudo em questão, se o sucesso for a reprovação de um aluno, e para um determinado vetor de características “*X”* dos alunos, a resposta média segundo o modelo for 0,7, isto significará que a probabilidade de reprovação desse aluno para o vetor de características “*X”* será de 70%.

2 - Objetivos

Gerais:

- Analisar o rendimento acadêmico dos alunos que cursaram as disciplinas Probabilidade e Estatística, Estatística Aplicada e Bioestatística durante o período de 1/2004 a 2/2008 com base em dados extraídos do Sistema de Informação de Graduação (SIGRA).

Específicos:

- Estudar as técnicas de análise de dados categorizados: Testes e Regressão Logística;
- Fazer a análise descritiva das variáveis do banco de dados;
- Verificar os índices de aprovação e reprovação, trancamentos e abandonos;
- Verificar se existe diferença de rendimento entre alunos de diferentes cursos, turnos, professor e outras variáveis de interesse;
- Identificar fatores associados ao rendimento acadêmico dos estudantes nas disciplinas em estudo, com base em dados cadastrais e de histórico escolar.

3 - Referencial Teórico

Ao estudar o rendimento de alunos onde as variáveis explicativas são em sua maioria qualitativas, é necessário abordar técnicas como a razão de chances e a regressão logística.

3.1 – Razão de Chances (Odds Ratio)

Primeiramente, define-se como chance a razão entre a probabilidade de ocorrência de um evento e a probabilidade de não ocorrência desse evento. Seja, por exemplo, X e Y variáveis que representam o sexo e aprovação em um certa disciplina, respectivamente. A representação tabular desta será:

$x \setminus y$	Sim	Não	Total
Masculino	π_{11}	π_{12}	π_1
Feminino	π_{21}	π_{22}	π_2

Razão de chances (Odds Ratio) é uma medida de associação usada para tabelas de contingência 2×2 e será fundamental para o modelo de regressão logística. A probabilidade de sucesso do nível 1 de uma variável categórica X pode ser representada por π_1, e as chances ($odds_1$) de sucesso são definidas como:

$$odds_1 = \frac{\pi_1}{1 - \pi_1} \tag{1}$$

De forma análoga, π_2 representa a probabilidade de sucesso do nível 2 e suas chances ($odds_2$) é definida como:

$$odds_2 = \frac{\pi_2}{1-\pi_2}.$$

As chances têm valores não negativos os quais são maiores que um quando o sucesso é mais provável de ocorrer.

Com o *odds* 1 e 2, foram apresentados as probabilidades de sucesso dos níveis 1 e 2 da variável X. Para qualquer nível da variável X, a probabilidade de sucesso é uma função das chances representada por:

$$\pi = \frac{odds}{odds + 1} \tag{2}$$

Quando as distribuições condicionais dos níveis da variável categórica X são idênticas, ou seja, $\pi_1 = \pi_2$ isso implica em $odds_1 = odds_2$, então as variáveis X e Y da tabela de contingencia 2×2 são independentes. A razão de chances dos dois níveis de X, mais conhecida como "*odds ratio*", é expressa por:

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{3}$$

Sendo que:

$\theta = 1$: X e Y são independentes e quanto mais distantes de um, maior a associação entre X e Y;

$0 < \theta < 1$: as chances de sucesso do nível 2 são maiores que do nível 1;

$1 < \theta < \infty$: as chances de sucesso do nível 1 são maiores que do nível 2.

A razão de chances possui uma importante propriedade, que compreende no fato que seus valores não dependem da orientação da tabela de contingência (qual variável assume a posição de linha e qual assume a de coluna), ou seja, a razão de chances (odds ratio) trata as variáveis simetricamente.

Quando ambas as variáveis são respostas, a razão de chances pode ser definida através das probabilidades conjuntas:

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \quad (4)$$

A partir dessa expressão, nota-se que θ é a razão do produto cruzado das diagonais da tabela 2×2, motivo pelo qual a razão das chances também é chamada de razão do produto cruzado.

A razão de chances para uma amostra tem a seguinte expressão:

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \quad (5)$$

Em pequenas amostras, a distribuição de $\hat{\theta}$ é muito assimétrica. Para contornar tal inconveniente, utiliza-se a transformação logarítmica para obter simetria em torno do zero, aproximando esta para a distribuição normal. Devido a esse fato, a estimativa da média é $\log(\theta)$ e o desvio padrão assintótico (ASE) é expressado por:

$$ASE\left[\log(\hat{\theta})\right] = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \quad (6)$$

O intervalo de confiança é construído para a estimativa $\log(\theta)$ e posteriormente aplica-se a transformação exponencial (transformação de volta) nos valores finais para a obtenção do intervalo de confiança para θ. Então, o intervalo de confiança (1-α) para o $\log(\theta)$ será:

$$IC((1-\alpha), \log(\theta)) = \log(\hat{\theta}) \pm z_{(\alpha/2)} ASE\left[\log(\hat{\theta})\right] \quad (7)$$

Caso haja caselas nulas ou valores muito pequenos nas tabelas de contingência, deve-se fazer uma pequena alteração no estimador da razão de chances para não se obter resultados indesejados como 0 ou ∞, neste caso a alteração será:

$$\tilde{\theta} = \frac{(n_{11} + 0{,}5) * (n_{22} + 0{,}5)}{(n_{12} + 0{,}5) * (n_{21} + 0{,}5)} \quad (8)$$

Na fórmula do $ASE\left[\log(\hat{\theta})\right]$, basta trocar $n_{ij}\ por\ (n_{ij} + 0{,}5)$.

Toda a teoria apresentada anteriormente refere-se a tabelas 2×2, porém a razão de chances pode ser estendida para tabelas de contingência com dimensões maiores. Para isso, basta escolher uma categoria como referência e comparar as demais com ela, obtendo várias comparações duas a duas em relação ao nível de referência adotado.

3.2 - Regressão Logística

Em pesquisas demográficas, há o interesse de saber se a chance de uma família estar ou não abaixo da linha da pobreza está relacionada com fatores como número de crianças, número de pessoas do sexo feminino e do masculino, número de aposentados, máximo de anos de estudo entre os membros da família e renda mensal familiar.

No que diz respeito à medicina, pode-se determinar fatores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos, podendo gerar um modelo que dirá a probabilidade de um indivíduo ter câncer dado suas características estudadas.

Em todos os exemplos acima, a regressão logística é bem aplicável, pois o modelo de regressão logística é utilizado quando a variável resposta é qualitativa, com dois resultados possíveis ou mais, podendo ser estendido quando a variável resposta possui mais de duas respostas, como por exemplo, se a pressão sanguínea é alta, média ou baixa.

Um modelo de regressão linear é um método para se estimar a reposta média de uma variável resposta (y) através de variáveis explicativas (x) observadas. O termo linear sugere que a relação entre a variável resposta e a variável explicativa pode ser descrita através de uma linha. Na maioria dos casos a variável resposta é contínua, em outros, ela pode ser dicotômica(binária) ou politômica.

Quando se trata do modelo de regressão linear logístico, a variável respostas se comporta de forma específica, sendo esta binária, ou seja, 0 e 1,

ou até mesmo com três ou mais categorias, como por exemplo se a pressão sanguínea é baixa, média ou alta. No caso de uma resposta binária, esta pode ser representada de diversas formas, como por exemplo: Sim ou não, Sobrevive ou não, compra ou não compra, dentre outros. O modelo logístico pode ser utilizado para analisar dados observacionais ou experimentais em um delineamento completamente casualizado.

3.2.1 - Modelo de Regressão para variável resposta binária

Quando a variável resposta é binária, a forma da função da variável resposta será sempre curvilínea. Uma propriedade importante da função logística é que ela pode ser linearizada.

O modelo de regressão linear simples geral é:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (9)$$

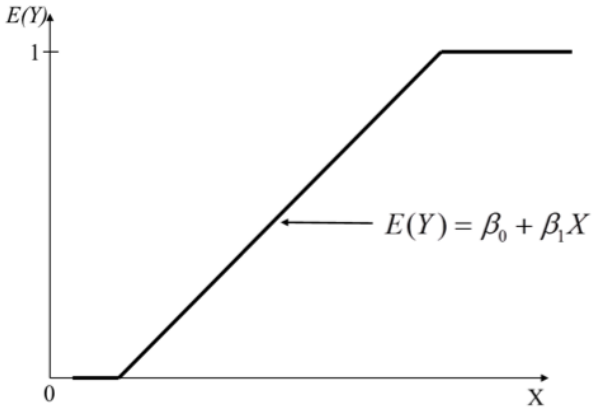
sendo que a variável resposta Y_i terá resultados binários. Considerando que Y_i seja uma Bernoulli com a seguinte distribuição de probabilidade:

$\boldsymbol{Y_i}$	Probabilidade
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Pela definição de média de uma distribuição discreta, a média da Bernoulli será:

$$E(Y_i) = \sum y_i P(Y_i = y_i) = 1(\pi_i) + 0(1 - \pi_i) = (\pi_i) = P(Y_i = 1)$$

Uma possível representação gráfica para o modelo logístico (9) poderia ser:



Observando este, porém, nota-se que como a resposta está entre “0” e “1”, sua interpretação final denota a probabilidade do sucesso ocorrer, como por exemplo, se o sucesso for a reprovação de um aluno, e para um determinado vetor de características “X” dos alunos, a resposta média segundo o modelo for 0,7, isto significará que a probabilidade de reprovação desse aluno para o vetor de características “X” é de 70%.

Existem três diferenças principais entre o modelo de regressão simples e o logístico:

a) Os erros não possuem distribuição normal;

O erro será igual a $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - E(Y_i) = Y_i - \pi_i$, podendo assumir dois valores:

- Quando $Y_i = 1$, $\varepsilon_i = 1 - (\beta_0 + \beta_1 X_i)$
- Quando $Y_i = 0$, $\varepsilon_i = -\beta_0 - \beta_1 X_i$

Logo, a suposição de que os erros são normalmente distribuídos não é válida.

b) Variância dos erros não é constante;

Como Y_i tem distribuição Bernoulli, sua variância é $\pi_i(1 - \pi_i)$, então a variância do erro ($\varepsilon_i = Y_i - \pi_i$, sendo π_i uma constante) será:

$$\sigma^2\{\varepsilon_i\} = \sigma^2\{Y_i\} + \sigma^2\{\pi_i\} = \sigma^2\{Y_i\} = \pi_i(1 - \pi_i)$$

onde $\pi_i(1-\pi_i) = (\beta_0 + \beta_1 X_i) \times (1 - \beta_0 - \beta_1 X_i)$, tem-se então que esta expressão depende de X_i. Assim, a variância do erro difere para os diferentes níveis de X.

c) Restrição na função resposta.

A resposta média $E(Y)$ representa a probabilidade de o sucesso ocorrer, logo $0 \leq E(Y) = \pi \leq 1$.

Com variáveis dependentes binárias, a função reposta adquire formato de *S* assintótico de 0 e 1. As funções respostas adequadas para tais características são:

- *Probit:* Assume que a distribuição dos erros é normal;
- Logística: Assume que os erros possuem distribuição logística, possuindo caldas mais pesadas;
- Complemento log-log: Usada quando a distribuição dos erros é assimétrica.

3.2.2 - Modelo de Regressão Logística Simples

O modelo logístico será semelhante a (9), sendo composto pela soma da média da variável resposta com um erro aleatório, ou seja:

$$Y_i = E(Y_i) + \varepsilon_i \tag{10}$$

sendo que Y_i tem distribuição de Bernoulli com parâmetro $E(Y_i) = \pi$. Visto que a distribuição dos erros depende da distribuição de Bernoulli da variável Y_i, é preferível atribuir-lhes a distribuição logística. A densidade de uma variável aleatória logística com média "0" e desvio padrão $\pi/\sqrt{3}$ será:

$$f_L(\varepsilon_L) = \frac{\exp(\varepsilon_L)}{[1 + \exp(\varepsilon_L)]^2} \tag{11}$$

Seja Y_i^c uma variável aleatória contínua, sendo $Y_i^c = \beta_0^c + \beta_1^c x_i + \varepsilon_i^c$. Dicotomizando Y_i^c de forma se obter uma variável binária, tem-se:

$$\begin{cases} Y_i = 1, & se\ Y_i^c \leq \mathrm{A}, \quad \mathrm{A} \in \mathbb{R} \\ Y_i = 0, & se\ Y_i^c > \mathrm{A}, \quad \mathrm{A} \in \mathbb{R} \end{cases}$$

Supondo ε_i^c, com distribuição logística com média 0 e desvio padrão σ_c. Assim a média da variável dicotomizada será:

$$\pi_i = P(Y_i = 1) = P(Y_i^c \leq A) = P(\beta_0^c + \beta_1^c x_i + \varepsilon_i^c \leq A)$$

Isolando ε_i^c: $$P(\varepsilon_i^c \leq A - \beta_0^c - \beta_1^c x_i)$$

Dividindo por σ_c:

$$P\left(\frac{\varepsilon_i^c}{\sigma_c} \leq \frac{A-\beta_0^c-\beta_1^c x_i}{\sigma_c}\right) = P\left(\frac{\varepsilon_i^c}{\sigma_c} \leq \beta_0^* + \beta_1^* x_i\right) \quad (12)$$

Sendo $\beta_0^* = \frac{A-\beta_0^c}{\sigma_c}$ e $\beta_1^* = \frac{\beta_1^c}{\sigma_c}$. Multiplicando ambos os lados da inequação (12) por $\pi/\sqrt{3}$ a probabilidade não será afetada, então:

$$P\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i^c}{\sigma_c} \leq \frac{\pi}{\sqrt{3}}\beta_0^* + \frac{\pi}{\sqrt{3}}\beta_1^* x_i\right) = P(\varepsilon_L \leq \beta_0 + \beta_1 x_i) = F_L(\varepsilon_L)$$

Logo tem-se que a distribuição acumulada será:

$$F_L(\varepsilon_L) = \frac{\exp(\varepsilon_L)}{1 + \exp(\varepsilon_L)} \quad (13)$$

Em suma, a função resposta média logística, π_i, será igual a equação (13), podendo ser expressada como $\pi_i = F_L(\varepsilon_L) = [1 - \exp(-\beta_0 - \beta_1 x_i)]^{-1}$. Aplicando a inversa nesta última expressão tem-se que:

$$\eta_i = F_L^{-1}(\pi_i) = log_e\left(\frac{\pi_i}{1-\pi_i}\right) \quad (14)$$

Sendo que (14) é chamado de transformação *logit* para a probabilidade π_i, conhecida também como preditor linear da média.

A razão $\pi_i/(1-\pi_i)$ da fórmula (14) é chamada de razão de chances (*odds ration*). Valendo frisar que a chance de um evento ocorrer é a razão entre a probabilidade de ocorrência desse evento e a probabilidade de não ocorrência desse evento.

3.2.2.1 - Estimação dos Parâmetros do Modelo

O método de estimação dos parâmetros da regressão logística é o Método de Máxima Verossimilhança (EMV). Este método de estimação procura obter o valor do parâmetro a ser estimado que maximiza a probabilidade de obter a amostra observada.

Sendo Y_i uma variável aleatória Bernoulli, sua função de densidade de probabilidade será:

$$P(Y_i = y_i) = {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i} \tag{15}$$

A função de probabilidade conjunta será o produtório da função (15), visto que os ${Y_i}'s$ são independentes.

$$L(\pi, y) = \prod_{i=1}^{n} {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i} \tag{16}$$

Aplicando *ln* (logaritmo neperiano) na função $L(\pi, y)$tem-se:

$$\ln\{L(\pi, y)\} = \sum_{i=1}^{n} Y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right) + \sum_{i=1}^{n} \ln(1 - \pi_i) \tag{17}$$

Derivando esta ultima expressão (17) em relação a π_i, tem-se :

$$1 - \pi_i = [1 + \exp(\beta_o + \beta_1 X_i)]^{-1}$$

Isolando π_i, tem-se que o estimador de máxima verossimilhança pra π_i será:

$$\hat{\pi}_i = \frac{\exp(b_o + b_1 X_i)}{1 + \exp(b_o + b_1 X_i)} \tag{18}$$

Os valores de b_o e b_1 são estimativas de β_o e β_1, que são obtidas através de métodos computacionais ou procedimentos numéricos. Quanto as variáveis X, estas podem ser quantitativas ou qualitativas (sendo estas

variáveis indicadoras), como mencionado anteriormente, e segundo Neter, esta flexibilidade torna o modelo logístico bastante interessante.

3.2.2.2 - Interpretação dos parâmetros do modelo

O valor da resposta da função logito ajustada em $X = x_j$ será o inverso da função (18), sendo está igual a :

$$\hat{\pi}'_i(x_j) = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = b_0 + b_1 x_j \qquad (19)$$

Na equação (19), b_0 e b_1 representam a estimativa calculada pela amostra para β_o e β_1 respectivamente. O razão de $\hat{\pi}_i$sobre $1-\hat{\pi}_i$ dentro do *ln* representa a razão de chances, seja esta denominada como *odds₁*. O valor de b_0 representará o intercepto da função, ou seja, será o ponto onde a função $\hat{\pi}'$ irá começar quando $X = 0$. O valor de b_1 representará a taxa de acréscimo à função $\hat{\pi}'$ quando acrescentada uma unidade em X. Uma outra abordagem seria dizer que b_1 será a diferença entre $\hat{\pi}'(x_j + 1)$ e $\hat{\pi}'(x_j)$, ou seja:

$$\hat{\pi}'(x_j + 1) - \hat{\pi}'(x_j) = b_1$$

$$\ln(odds\ 2) - \ln(odds\ 1) = \ln\left(\frac{odds\ 2}{odds\ 1}\right) = b_1 \qquad (20)$$

Assim

$$\widehat{OR} = {}^{odds\ 2}/_{odds\ 1} = \exp(b_1),$$

ou seja, $\widehat{OR}$ representa a *odds ratio* ou razão de chances. Esta última medida fornecerá a variação de chances para cada unidade acrescentada em X. Contudo, se a variável X for qualitativa (sexo, por exemplo), então $\exp(b_1)$ representará a variação da chance de uma de suas categorias em relação a outra (categoria de referência).

3.2.3 - Modelo de Regressão Logística Múltipla

O modelo Múltiplo compreende uma extensão do modelo de regressão logística simples adicionando mais variáveis explicativas ao modelo. Tem-se que o modelo simples será igual a (19), já o modelo multivariado será:

$$\hat{\pi}' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1} \tag{21}$$

$$\hat{\pi}_i' = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1} \tag{22}$$

Reescrevendo os modelos acima de forma matricial, tem-se:

$$\hat{\pi}_i' = \acute{X}\beta = \begin{bmatrix} 1 \\ x_1 \\ . \\ . \\ . \\ x_{p-1} \end{bmatrix} \times \begin{bmatrix} \beta_0 & \beta_1 \dots & \beta_{p-1} \end{bmatrix}$$

A transformação *logito* para o modelo múltiplo pode ser entendida como uma extensão da obtida pelo modelo simples, fornecendo:

$$E(Y) = \frac{1}{1 + \exp(-x'\beta)} \tag{23}$$

- **Inferências sobre o parâmetro de regressão**

Para os procedimentos inferenciais que serão apresentados é preciso ter amostras grandes para que o EMV da regressão logística tenha distribuição aproximadamente normal.

Antes da análise dos parâmetros de regressão se faz necessário analisar se estes são relevantes ao modelo ajustado. Para testar se estes parâmetros são significativos pode ser utilizado o teste de Wald. O teste de Wald consiste em dizer quais variáveis são estatisticamente relevantes para o

modelo e se existe diferença na grandeza entre elas. Assim as hipóteses do teste para um β_k serão:

$$\begin{cases} H_0: \ \beta_k = 0 \\ H_0: \ \beta_k \neq 0 \end{cases}$$

A estatística do teste será:

$$Z^* = \frac{b_k}{S\{b_k\}} \sim N(0,1) \qquad (24)$$

Para um determinado nível de significância α, a regra de decisão para concluir H_0 será $|Z^*| \leq Z_{(1-\alpha/2)}$, o contrário desta implicará a rejeição de H_0.

Para testes unilaterais basta utilizar a regra de decisão unilateral com suas devidas modificações. Frequentemente, a estatística $(Z^*)^2$ é utilizada no lugar de Z^*, esta nova estatística terá distribuição Qui-Quadrado com 1 grau de liberdade.

- **Intervalo de Confiança para um único β_k**

Antes de analisar os intervalos de confiança, faz-se necessário entender o seu real significado. Com α de significância e "n" amostras independentes, um intervalo de $1-\alpha$ de confiança nos diz que $(1-\alpha)\%$ das médias dessas amostras, por exemplo, estarão dentro do intervalo especificado.

O intervalo de confiança para β_k com $1-\alpha$ de confiança é dado por:

$$\beta_k \ \pm \ Z_{(1-\alpha/2)} S\{\beta_k\} \qquad (25)$$

Quanto ao *odds ratio,* seu intervalo de confiança correspondente será:

$$\exp[\beta_k \ \pm Z_{(1-\alpha/2)} S\{\beta_k\}\,] \qquad (26)$$

O procedimento de Bonferroni oferece intervalos de confiança simultâneos para *"g"* parâmetros da regressão logística. Se estes "g" parâmetros são estimados, Bonferroni oferece o seguinte intervalo de confiança.

$$b_k \pm B.S\{\beta_k\} \tag{27}$$

sendo $B = Z_{(1-\alpha/2g)}$.

- **Teste da Razão de Verossimilhança**

Análogo ao teste linear geral para modelos lineares, o Teste da Razão de Verossimilhança é baseado na comparação entre os modelos reduzido e completo, tendo como objetivo verificar se alguns β_k's são iguais a zero, sendo necessário tamanhos grandes de amostras. O Teste da Razão de Verossimilhança envolve a razão da função de verossimilhança do modelo logístico completo e a função de verossimilhança do modelo sob H_0. As hipóteses dos testes serão:

$$\begin{cases} H_0: \quad \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ H_a: pelo\ menos\ um\ \beta_k \neq 0\ de\ H_0 \end{cases}$$

O modelo completo(*full*) será:

$$\pi_F = [1 + \exp -\{\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}\}]^{-1} \tag{28}$$

O modelo reduzido será (para $q < p$):

$$\pi_R = [1 + \exp -\{\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{q-1}\}]^{-1} \tag{29}$$

A estatística do teste será denotada por G^2, sendo esta igual a:

$$G^2 = -2[\ln(R) - \ln(F)] = -2\left[\ln\left(\frac{R}{F}\right)\right] \tag{30}$$

A hipótese nula será rejeitada se G^2 pertencer a região crítica, ou seja, se $G^2 > \chi^2_{(1-\alpha;p-q)}$.

3.2.4 - Métodos de Seleção do Modelo

Existem diversos métodos de seleção de variáveis para obter o modelo final. Para isso, é preciso estabelecer critérios de seleção e, em regressão

logística, os utilizados são: -2logRV, AIC e SBC, estes últimos também utilizados em regressão linear. Tais critérios se fazem necessários visto que em um modelo com "p" variáveis preditoras, 2^p modelos de regressão podem ser gerados.

O Teste de Razão de Verossimilhança consiste na comparação entre os modelos reduzido e completo, tendo como objetivo verificar se alguns β_k's são iguais a zero, sendo necessário tamanhos grandes de amostras.

Já os critérios de AICp e SBCp consistem em:

$$AICp = -2\ln[L(b)] + 2p \tag{31}$$

$$SBCp = -2\ln[L(b)] + p\ln(n) \tag{32}$$

sendo $\ln[L(b)] = \sum_{i=1}^{n} Y_i(X_i'\beta) - \sum_{i=1}^{n} \ln[1 + \exp(X_i'\beta)]$. O melhor modelo ajustado apresentará menor AIC e SBC. Note que os fatores "$2p$" e "$p\ln(n)$" adicionados a AIC e SBC respectivamente são penalidades baseadas no acréscimo de parâmetros no modelo, o que dificulta sua interpretação.

Porém, quando há muitas variáveis, os procedimentos acima se tornam inviáveis devido à quantidade de valores gerados para os diversos modelos. Sendo assim, métodos de seleção automática de variáveis para a construção de modelos são comumente utilizados.

Um destes procedimentos de seleção automática de variáveis é o *Stepwise*, que consiste em construir iterativamente uma sequência de modelos de regressão pela adição ou remoção de variáveis em cada etapa. Para modelos logísticos, a regra de decisão para a entrada e saída de variáveis no modelo é baseada na estatística Z^* de *Wald* para o parâmetro β_k e seu p-valor.

O procedimento de *Stepwise* é exclusivo, uma vez que uma variável preditora sai do modelo, este método desconsidera qualquer outra combinação com tal variável, portanto exclui modelos que poderiam ser mais adequados. Porém os critérios de AICp e de SBCp podem ajudar na indicação de outros modelos, por isso faz-se necessária uma análise apurada (testar adequabilidade e ajustamento) de cada modelo gerado para ver qual melhor se encaixa com o problema real.

3.2.4.1 – Testes para adequabilidade de ajustamento

Depois de se obter o modelo através dos métodos de seleção mencionados, deve-se verificar sua adequabilidade através de propriedades básicas da função resposta estimada.

- **Teste Qui-quadrado de Pearson**

O teste Qui-quadrado de Pearson necessita de duas suposições para seu cálculo, que é a de independência das observações e de réplicas das mesmas. Além disto, o teste Qui-quadrado de Pearson não detecta pequenos desvios da função de regressão logística. As hipóteses deste teste serão:

$$\begin{cases} H_0: E(y) = [1 + \exp(-X'\beta)]^{-1} \\ H_a: E(y) \neq [1 + \exp(-X'\beta)]^{-1} \end{cases}$$

A estatística do teste será:

$$\chi^2 = \sum_{j=1}^{c} \sum_{k=0}^{l} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \tag{33}$$

sendo que:

$X_j =$ Valor de X para o nível j.

$Y_{ij} =$ Valor da variável binária Y na i-ésima replicação do j-ésimo nível de X.

$n_j =$Número de replicações do j-ésimo nível de X.

$c =$Número de combinações distintas das variáveis preditoras.

$$\sum_{i=1}^{n_j} Y_{ij} = Y_{.j}$$

$O_{j1} = \sum_{i=1}^{n_j} Y_{ij} = Y_{.j} =$ Número de casos (replicações) do j-ésimo nível com a variável resposta igual a um.

$O_{j0} = \sum(1 - Y_{ij}) = n_j - Y_{.j} = n_j - O_{j1} =$ Número de casos (replicações) do j-ésimo nível com variável resposta igual a zero.

$E_{j1} = n_j \times \hat{\pi}_j$

$E_{j0} = n_j \times (1 - \hat{\pi}_j)$

A hipótese nula será rejeitada quando $\chi^2 > \chi^2_{(1-\alpha;c-p)}$.

- **Teste Deviance**

Deviance representa duas vezes a diferença do *log* do modelo saturado menos do *log* do modelo a ser testado, ou seja:

$$D = 2(\hat{l}_n - \hat{l}_p) \qquad (34)$$

O teste de *Deviance* para a adequabilidade de ajustamento é análogo ao teste de falta de ajustamento para modelos de regressão lineares e utiliza a mesma notação que o teste Qui-quadrado de Pearson. Portanto, nota-se que tal teste também necessita de repetições de níveis da variável preditora.

Como no teste linear geral, o Teste *Deviance* utiliza o modelo reduzido (29) e o modelo completo (28) para realizar o teste da razão da verossimilhança (30). Para a realização desse teste é necessário os valores que maximizam as verossimilhanças dos modelos reduzido e completo denotados por $L(R)$ e $L(F)$ respectivamente.

A estimativa de $L(R)$ é obtida pelo ajuste do modelo reduzido e $L(F)$ pela proporção amostral $p_j = \frac{Y_{.j}}{n_j}$ para $j = 1, \dots, c$. Dessa forma a estatística do teste é expressa por:

$$G^2 = -2[\ln(R) - \ln(F)] = -2\left[\ln\left(\frac{R}{F}\right)\right]$$

$$G^2 = -2\sum_{j=1}^{c}\left[Y_{.j}\ln\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{.j})\ln\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right]$$

$$G^2 = Dev(X_0, X_1, \dots, X_{p-1}) \qquad (35)$$

As hipóteses deste teste serão:

$$\begin{cases} H_0: E(y) = [1 + \exp(-X'\beta)]^{-1} \\ H_a: E(y) \neq [1 + \exp(-X'\beta)]^{-1} \end{cases}$$

Tendo visto as hipóteses acima, haverá evidências para rejeitar a hipótese nula se $Dev(X_0, X_1, \dots, X_{p-1}) > \chi^2_{(1-\alpha;c-p)}$.

- **Teste de Ajustamento de Hosmer – Lameshow**

A suposição de repetições nos níveis das variáveis independentes, como visto nos dois testes anteriores (Teste Qui-quadrado de Pearson e Teste de Deviance), não é necessária para o teste de Hosmer – Lameshow, o qual pode ser usado tanto para um conjunto de dados com poucas repetições quanto sem repetições.

O procedimento desse teste consiste no agrupamento dos dados em classes que possuam valores ajustados ($\hat{\pi}_i$) próximos, e que contenham o mesmo número de observações aproximadamente. A estatística do teste é calculada pela estatística Qui-quadrado de Pearson (33) a partir dos valores observados e esperados de uma tabela $C \times 2$, sendo esta uma estatística com $c - 2$ graus de liberdade.

As classes deste teste podem ser obtidas baseadas em duas maneiras diferentes: nos percentis das probabilidades estimadas e em valores fixados das probabilidades estimadas. A primeira forma grupos pelos pares de frequências observadas e estimadas com $^{n}/_{c}$ menores valores de probabilidades estimadas e a segunda, pelos pares de frequências estimadas e observadas com valores de probabilidades estimadas entre pontos de corte adjacentes definidos nos valores $^{k}/_{c}$, $k = 1, \dots, (c-1)$.

3.2.5 - Diagnósticos da Regressão Logística

Após investigar qual o modelo que melhor se ajusta aos dados, é imprescindível analisar os resíduos (diferenças entre os dados observados e os valores ajustados) produzidos por esse modelo. Tais resíduos não devem apresentar padrões bem definidos, já que se deseja que o modelo capte toda a estrutura de dependência da variável Y modelada.

Para a regressão logística, a análise de resíduos é mais complicada que a da regressão linear, já que Y_i tem distribuição Bernoulli. Seja Y_i uma variável binária que assume valores "0" ou "1", o i-ésimo resíduo apresentará um dos seguintes valores:

$$e_i = \begin{cases} 1 - \hat{\pi}_j, se\ Y_i = 1 \\ -\hat{\pi}_j, se\ Y_i = 0 \end{cases}$$

Nota-se que os resíduos não terão distribuição normal.

3.2.5.1 – Tipos de Resíduos

- **Resíduos de Pearson**

Os resíduos de Pearson são obtidos a partir da divisão do resíduo comum $(Y_i - \hat{\pi}_i)$ pelo erro padrão estimado de Y_i:

$$r_{pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}} \qquad (36)$$

É possível mostrar que a soma de quadrados dos resíduos de Pearson é numericamente igual a estatística do teste de Qui-quadrado de Pearson. Logo, o quadrado de cada resíduo de Pearson fornece a contribuição de cada resposta binária para o teste de adequabilidade.

- **Resíduos de Pearson Studentizados (RPS)**

Um procedimento melhor e bastante utilizado é chamado de Resíduos de Pearson Studentizados. Este divide os resíduos comuns pelo seu desvio padrão estimado que é, aproximadamente $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$, sendo h_{ii} é o i-ésimo elemento da diagonal da matriz chapéu estimada $n \times n$ para a regressão logística. A matriz chapéu satisfaz $\hat{\pi}' = \boldsymbol{HY}$ e sua expressão é dada por:

$$\boldsymbol{H} = \widehat{W}^{1/2}.X(X'.\widehat{W}.X)^{-1}.X'.\widehat{W}^{1/2} \tag{37}$$

sendo:

$\widehat{W} =$ Matriz diagonal $n \times n$ com os elementos $\hat{\pi}_i(1-\hat{\pi}_i)$.

$X =$ Matriz de delineamento $n \times p$.

Assim, os resíduos Studentizados de Pearson podem ser expressos por:

$$r_{spi} = \frac{r_{pi}}{\sqrt{1-h_{ii}}} \tag{38}$$

- **Resíduos Deviance**

Para dados binários, tem-se o seguinte modelo Deviance:

$$DEV(X_o, \dots, X_{p-1}) = -2\sum_{i=1}^{n}[Y_i \ln(\hat{\pi}_i) + (1-Y_i)\ln(1-\hat{\pi}_i)] \tag{39}$$

A partir desse, define-se o i-ésimo resíduo Deviance como:

$$dev_i = sinal\ (Y_i - \hat{\pi}_i)\sqrt{-2[Y_1 \ln(\hat{\pi}_i) + (1-Y_i)\ln(1-\hat{\pi}_i)]} \tag{40}$$

sendo que o "*sinal*" será positivo quando $Y_i > \hat{\pi}_i$ e negativo caso contrário. Assim:

$$\sum_{i=1}^{n}(dev_i)^2 = DEV(X_o, \dots, X_{p-1}) \tag{41}$$

Da expressão acima, nota-se que o quadrado de cada resíduo Deviance mede a contribuição de cada resposta binária para a estatística do Teste de Adequabilidade de Ajustamento Deviance (35).

3.2.5.2 - Gráficos de Resíduos

Os Gráficos de Resíduos são bastante utilizados para diagnosticar a inadequabilidade do modelo, homocedasticidade da variância e presença de *outliers*. No que se trata do modelo logístico, utilizam-se os gráficos de resíduos para apenas diagnosticar a inadequabilidade do modelo, visto que os resíduos logísticos não possuem variância constante (Seção 3.2.1).

3.2.5.3 – Detecção de observações influentes

Para a detecção de observações influentes, pode-se considerar a influência individual destas sob três aspectos da análise: Qui- quadrado de Pearson, Estatística Deviance e preditor linear ajustado $\hat{\pi}'_i$. Ambos utilizam o método da exclusão para analisar seu efeito nos resultados.

Para avaliar se um dado pode ser qualificado como influente ou não, é necessário ter um limite que possa dizer até onde a observação não é influente. A decisão é feita de forma subjetiva a partir dos gráficos: ΔX_i^2 e Δdev_i pela i-ésima observação ou pelo preditor linear estimado. Sendo:

$$\Delta X_i^2 = X^2 - X_{(i)}^2 \text{ e } \Delta dev_i = DEV - DEV_I \qquad (43)$$

Uma observação é considerada influente quando aparecem valores extremos como pontas (bem distantes do zero) no gráfico ΔX_i^2 pela i-ésima observação, ou quando nota-se a presença de *outliers* nos cantos superiores do gráfico Δdev_i pela i-ésima observação.

3.2.6 – Inferências sobre a resposta média

Para fazer inferência sobre a reposta média, é comum usar a notação matricial. Assim, as observações de X podem ser escritas como um vetor $p \times 1$, ou seja:

$$X_h = \begin{bmatrix} 1 \\ X_{h,1} \\ . \\ . \\ . \\ X_{h,p-1} \end{bmatrix}$$

Com o vetor X_h , estima-se $\hat{\pi}_h$ pontualmente como $\hat{\pi}_h = [1 + \exp(-X'_h b)^{-1}]$, sendo que "$b$" é o vetor dos parâmetros estimados. A partir da expressão de $\hat{\pi}_h$ e do fato que $\hat{\pi}'_h = X'_h b$, a expressão da estimativa pontual pode ser reescrita como:

$$\hat{\pi}_h = [1 + \exp(-\hat{\pi}'_h)]^{-1} \tag{44}$$

Para construir o intervalo de confiança é necessária a variância estimada aproximada de $\hat{\pi}'_h$, que é dada pela expressão:

$$S^2\{\hat{\pi}'_h\} = S^2\{X'_h b\} = X'_h S^2\{b\} X_h \tag{45}$$

sendo que $S^2\{b\}$ é a matriz de variâncias e covariâncias estimadas aproximadas dos parâmetros de regressão. Logo o intervalo de confiança para π'_h será:

$$\begin{cases} S = \hat{\pi}'_h + Z_{(1-\alpha/2)} S^2\{\hat{\pi}'_h\} \\ I = \hat{\pi}'_h - Z_{(1-\alpha/2)} S^2\{\hat{\pi}'_h\} \end{cases} \tag{46}$$

Finalmente, o intervalo de confiança da resposta média π_h será:

$$\begin{cases} Limite\ superior = [1 + \exp(-S)]^{-1} \\ Limite\ inferior = [1 + \exp(-I)]^{-1} \end{cases} \tag{47}$$

Note que os limites para π_h não são simétricos em torno da estimativa pontual, já que $\hat{\pi}_h$ não é uma função linear de $\hat{\pi}'_h$. O método de Bonferroni para intervalos de confiança simultâneo para várias respostas médias também pode ser empregado, basta substituir $Z_{(1-\alpha/2)}$ por $Z_{(1-\alpha/2g)}$ em "S" e "I", sendo que "g" é o número de intervalos de confiança simultâneos que deseja-se obter.

4 - Descrição do Problema

O Departamento de Estatística (EST) da Universidade de Brasília está vinculado ao Instituto de Ciências Exatas e foi criado em 20 de setembro de 1974. Atualmente, o Departamento é responsável pelo curso de Bacharelado em Estatística e pelo programa de Mestrado em Estatística. Além das disciplinas oferecidas para o Bacharelado e para o Mestrado, o Departamento de Estatística oferece aos demais cursos de graduação da UnB, três disciplinas com o conteúdo básico semelhante, variando em profundidade de acordo com a área (Vide nos anexos 1, 2 e 3 as ementas e os programas das disciplinas). As disciplinas são as seguintes:

- Probabilidade e Estatística (para área de Ciências Exatas e Economia)
- Estatística Aplicada (para área de Ciências Humanas)
- Bioestatística (para área de Ciências da Terra e da Vida)

A oferta destas disciplinas corresponde a pelo menos metade das turmas/disciplinas oferecidas pelo departamento por semestre e vem aumentando ao longo do tempo. Em particular, nos últimos semestres com a criação de novos cursos e aumento das vagas em vários dos cursos já existentes. Em 2º/2008 foram ofertadas seis turmas de Probabilidade e Estatística, nove turmas de Estatística Aplicada e duas de Bioestatística, totalizando dezessete turmas. No 2º/2011 esta oferta aumentou para vinte e oito turmas, sendo onze de Probabilidade e Estatística, 13 de Estatística Aplicada e quatro de Bioestatística.

Devido à grande demanda e variedade de cursos atendidos é fundamental estar atento ao rendimento dos estudantes nestas disciplinas, principalmente acompanhar os índices de reprovação e buscar entender as características dos reprovados e das disciplinas. Sendo assim, é importante um estudo aprofundado do rendimento dos estudantes das disciplinas Probabilidade e Estatística, Estatística Aplicada e Bioestatística.

5 – Metodologia

Neste trabalho foi utilizado o banco de dados extraído do Sistema de Informação Acadêmica (SIGRA), contendo informações sobre cada estudante que cursou uma das três disciplinas citadas outrora, além de características sócias demográficas fornecidas no registro ao ingressar na universidade, tais como: sexo, país de origem, naturalidade, disciplina cursada, menção obtida, porcentagem de faltas, professor, turno dentre outras. O arquivo contém 10.560 observações (alunos) abrangendo o primeiro semestre de 2004 até o segundo semestre de 2008. Vale frisar que este total de alunos (10.560) incluem as menções: SR, II, MI, MM, MS e SS.

Inicialmente, foi necessário um trabalho intenso e exaustivo de análise de consistência do banco de dados, incluindo desde a padronização das variáveis que possuíam erros de digitação até categorizar algumas delas, como por exemplo, "cidade onde o aluno mora"[1], "formas de ingresso na UNB"[2], devido à inviabilidade de análise diante da diversidade de respostas presentes.

Em seguida, fez-se a análise descritiva das variáveis a fim de conhecê-las melhor e posteriormente, geraram-se os modelos de interesse. Por último, foi feita a análise de resíduos com o intuito de verificar a existência de possíveis observações discrepantes.

[1] Consultar apêndice 3 para categorização segundo distância, formação das cidades e condições sócio-econômicas.
[2] Consultar o apêndice 2.

6 – Resultados

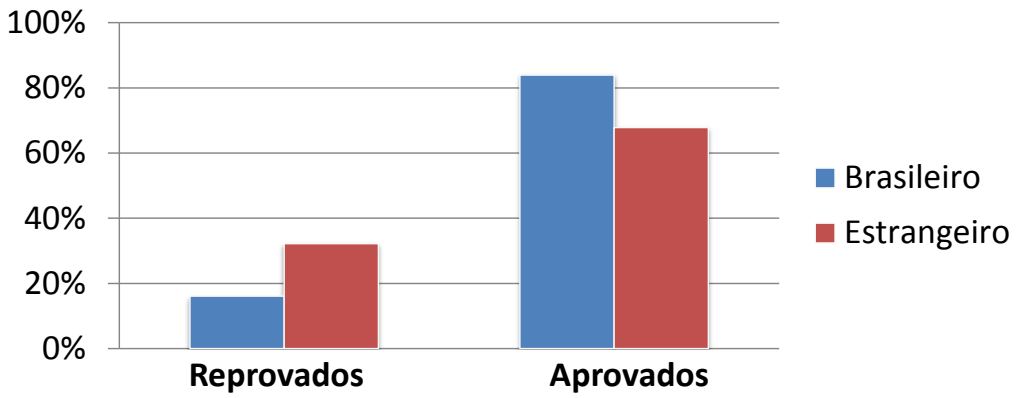
6.1 - Descrição dos Respondentes

Considerando como reprovação as menções iguais a “MI” e “II” (desconsiderando os 1388 alunos que obtiveram SR nas disciplinas), tem-se que de 35,56% de mulheres cerca de 12% reprovaram. O teste Qui-quadrado de independência revelou dependência entre o sexo e a reprovação($\alpha < 0{,}001$).

Quanto às disciplinas, cerca de 53,3% do total de alunos cursavam Estatística Aplicada, seguida por Probabilidade e Estatística (P.E.) com 30,85%. No que diz respeito à reprovação, a disciplina P.E. obteve maior índice, igual a 21,91%. As disciplinas E.A. e Bioestatística (BIO) tiveram índices de reprovação próximos a 14% (14,23% e 13,44%, respectivamente). O teste Qui-quadrado revelou que existe relação entre a reprovação e as disciplinas cursadas. Quanto ao turno em que a disciplina foi ministrada, cerca de 45,9% delas foram cursadas no turno matutino, e 28% no vespertino.

Quanto a nacionalidade dos alunos, cerca de 97,5% eram brasileiros, destes cerca de 16% reprovaram. 32% dos estrangeiros foram reprovados.

Gráfico 1: Índice de reprovação por nacionalidade no período de 2004 a 2008



Em relação ao curso de cada aluno, cerca de 11,8% cursavam Administração, 7,68% Ciências Contábeis e 22,9% cursavam alguma Engenharia.

Tabela 1: Porcentagem de Estudantes por Curso em Estatística Aplicada.

Curso	Porcentagem
Administração Diurno	11,77%
Administração Noturn	9,90%
Agronomia	0,23%
Antropologia	3,61%
Arquivologia	10,83%
Biblioteconomia	9,58%
Ciência Política	6,88%
Ciências Contábeis	14,14%
Ciências Sociais	4,90%
Computação	0,02%
Engenharia Civil	0,04%
Engenharia de Redes	0,02%
Engenharia Elétrica	0,02%
Engenharia Florestal	0,11%
Engenharia Mecânica	0,12%
Geografia	6,40%
Matemática	0,35%
Outros	7,23%
Relações Internacion	9,77%
Sociologia	4,10%

Tabela 2: Porcentagem de Estudantes por Curso em Bioestatística.

Cursos	Porcentagem
Agronomia	25,14%
Eng. Florestal	25,45%
Med. Veterinária	17,69%
Outros	31,72%

Tabela3: Porcentagem de Estudantes por Curso em Probabilidade e Estatística.

Curso	Porcentagem
Ciência da Computação	11,09%
Computação	11,55%
Eng. Civil	15,53%
Eng. Elétrica	12,85%
Eng. Mecatrônica	8,33%
Eng. Mecânica	11,46%
Eng. De Redes	8,78%
Matemática	12,37%
Outros	8,05%

Gráfico 2: Índice de reprovação por curso para estudantes que fazem Estatística Aplicada.

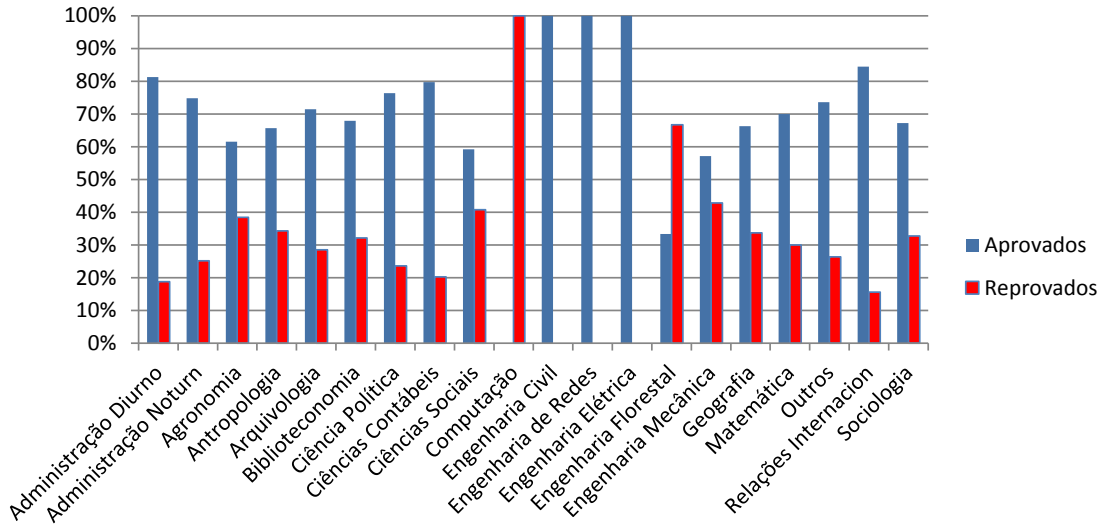


Gráfico 3: Índice de reprovação por curso para estudantes que fazem Bioestatística.

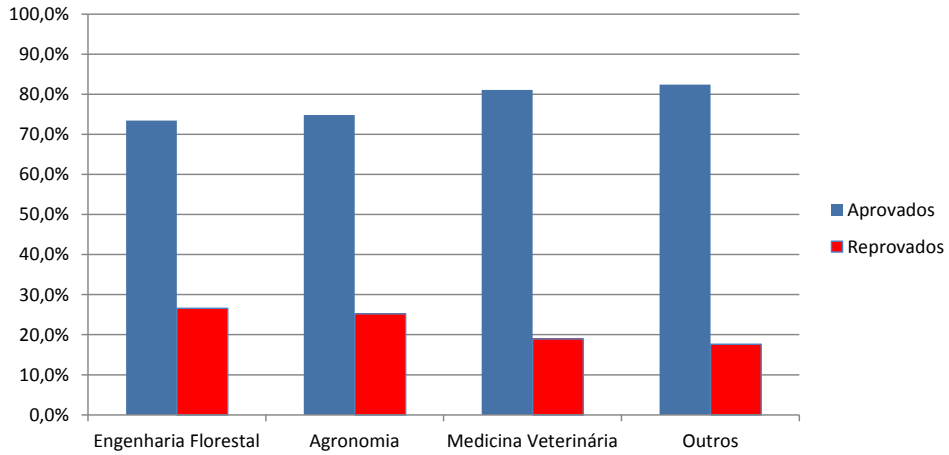
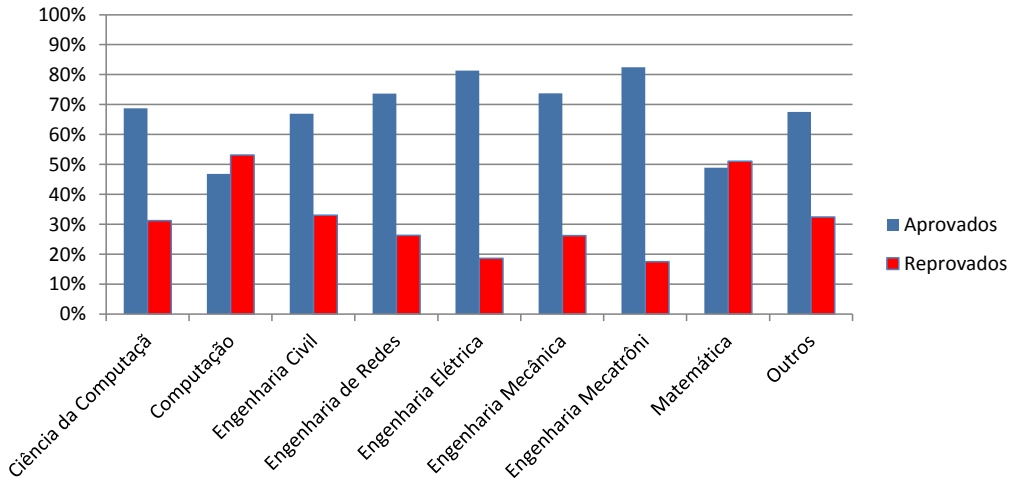


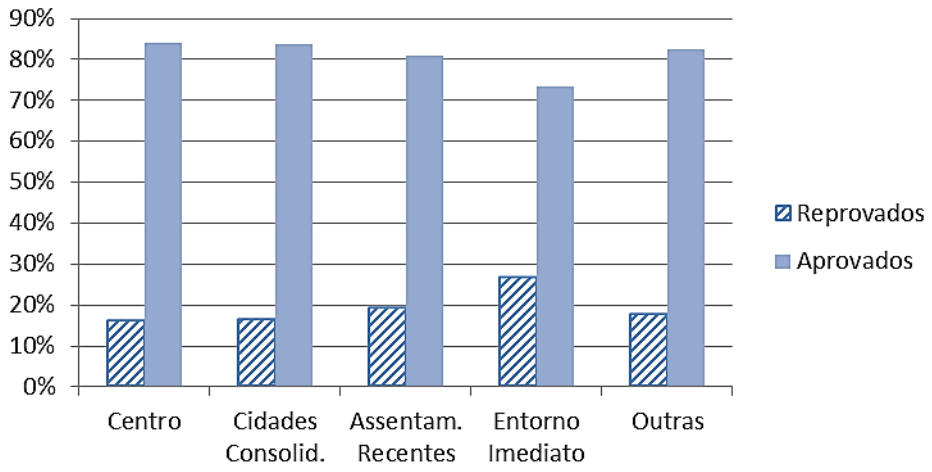
Gráfico 4: Índice de reprovação por curso para estudantes que fazem Probabilidade e Estatística.



Do total de alunos, 41,2% não cursaram a disciplina no fluxo, destes quase 22% reprovaram, e dos 48,5% que cursaram no fluxo, 13,14% reprovaram. O teste Qui-quadrado revelou que existe dependência entre a reprovação e o fato de o aluno ter cursado a disciplina no fluxo ou não.

Em relação ao local onde o aluno reside, aproximadamente 66,6% moram no centro do Distrito Federal, próximos ao Campus Darcy Ribeiro da UnB. Cerca de 20% moram em cidades consolidadas, 1,14% moram no entorno imediato e 10% moram em assentamentos recentes. Dos moradores próximos à UnB, cerca de 16% reprovaram e em relação aos moradores dos assentamentos recentes e do entorno imediato cerca de 10% e 26,7% reprovaram respectivamente.

Gráfico 5: Índice de reprovação por local de residência no período de 2004 a 2008

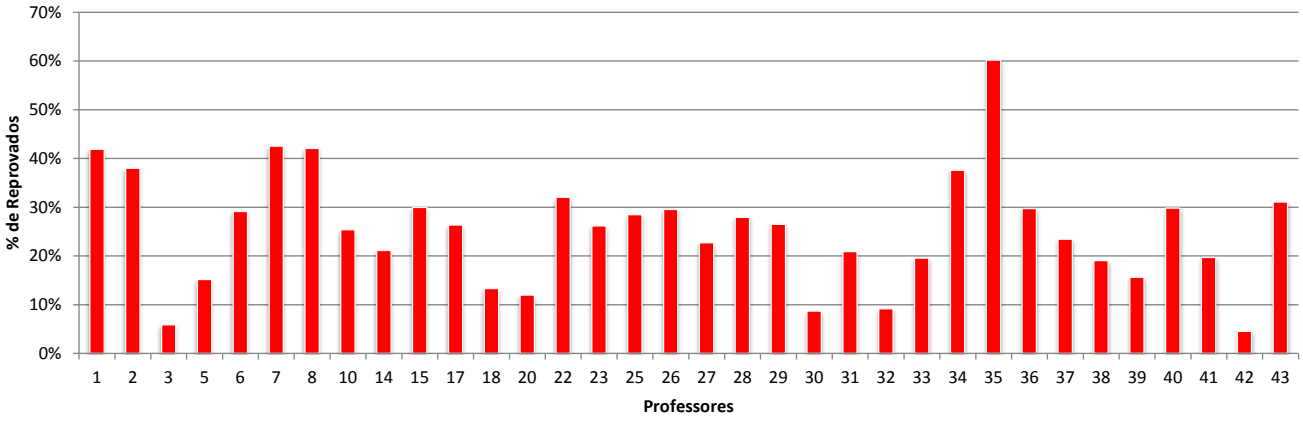


Dos cursos estudados, 57% eram de Bacharelado e 36,8% de cunho Profissional. Os cursos de licenciatura obtiveram o maior índice de reprovação, igual a 23,63%. Neste caso, o teste de independência Qui-quadrado revelou que existe uma dependência entre o tipo de curso (Bacharelado, Licenciatura, Profissional) e a reprovação.

Em relação ao número de vezes que fez a disciplina, cerca de 85% cursaram pela primeira vez, 11,4% pela segunda e 3,1% pela terceira vez. Dos alunos que cursaram pela primeira vez, 14,98% reprovaram e em relação aos que cursaram pela segunda vez, 25,36% reprovaram.

Cerca de 38 professores (do quadro regular ou não) ministraram algumas das disciplinas estudadas, destes, o professor de código igual a 35 obteve índice de reprovação igual a 38,8%, seguidamente, o professor de código 7 obteve 35,5% de reprovação. Vale frisar que estas reprovações não estão incluindo os abandonos (SR's). O professor que ministrou aula para mais alunos foi o de número 17, totalizando 7,5% do total de alunos.

Gráfico 6: Índice de reprovação por Professor no período de 2004 a 2008.



Fonte: Banco de Dados SIGRA

6.2 - Análise Descritiva

Uma breve análise descritiva sobre os índices de reprovação e abandono no período de 2004 a 2008 já inspira uma análise aprofundada sobre as características dos alunos que influenciam na reprovação.

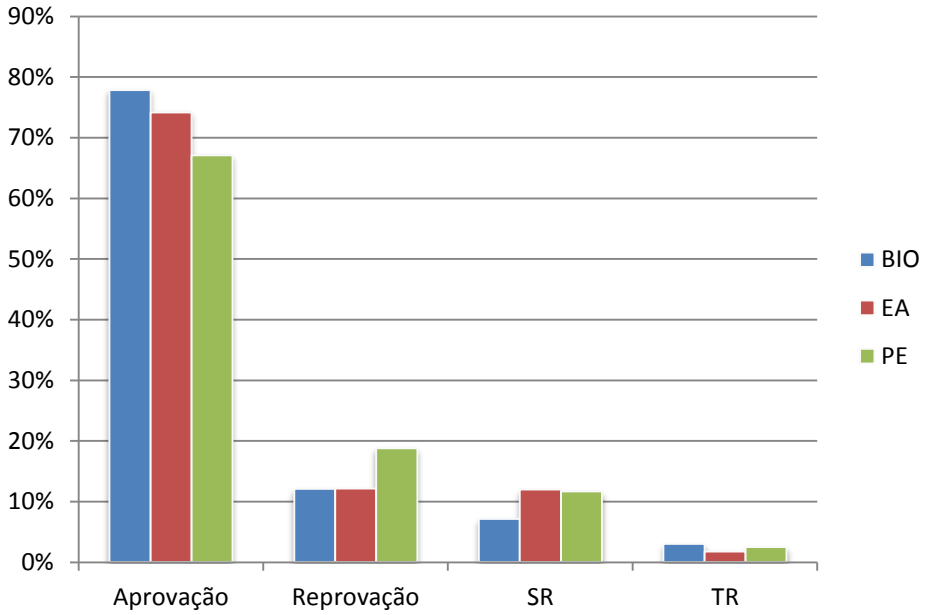
Tabela 4 – Rendimento dos estudantes segundo disciplina no período de 2004 a 2008

	BIO	EA	PE
Aprovação	76,70%	83,10%	73,90%
Reprovação	12,70%	13,50%	20,40%
SR	7,50%	2,40%	4,00%
TR	3,10%	1,00%	1,70%

Fonte: Banco de Dados SIGRA

Pelo Gráfico 7, tem-se que a disciplina Probabilidade e Estatística apresenta o maior índice de reprovação. Nota-se também que a disciplina Bioestatística apresentou maior índice de abandono. Quanto ao trancamento, a disciplina Estatística Aplicada obteve menor índice em relação às demais.

Gráfico 7: Rendimento dos estudantes segundo disciplina no período de 2004 a 2008.



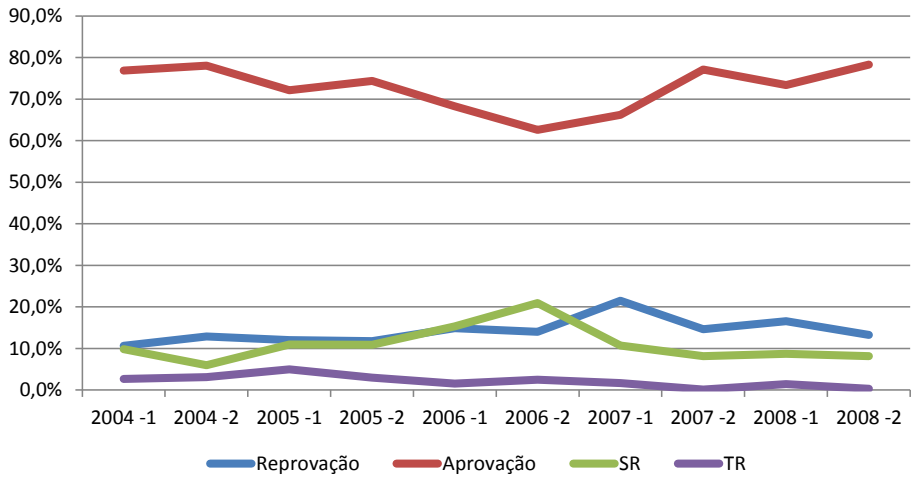
Fonte: Banco de Dados SIGRA

O gráfico 8 apresenta as porcentagens de aprovação, reprovação, SR e TR ao longo do tempo, e nota-se que a aprovação reduziu singelamente no período de referência.

Ao observar a linha de trancamento ao longo do tempo no gráfico 8, verifica-se que esta obteve pequenas variações próximas de zero, tendo como valor máximo 5%. A porcentagem de alunos que abandonaram as disciplinas ou obtiveram menções “sem rendimento” teve uma variabilidade razoável, chegando até superar a porcentagem de reprovação no 2º semestre de 2006, sendo este semestre o pico de SR’s.

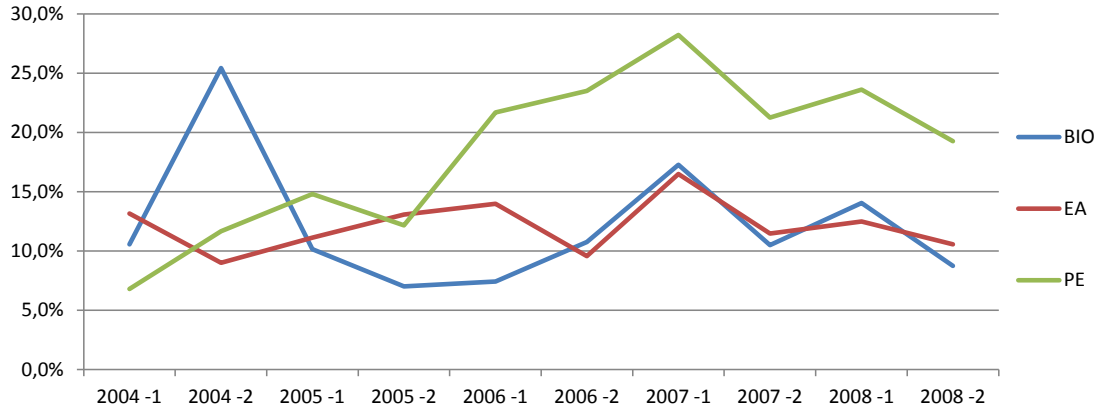
Ainda pelo gráfico 8, é possível visualizar que a reprovação aumentou suavemente ao longo do tempo estudado, obtendo seu ápice no primeiro semestre de 2007.

Gráfico 8: Rendimentos dos Estudantes no período de 2004 a 2008.



Fonte: Banco de Dados SIGRA

Gráfico 9: Índice de reprovação no período de 2004 a 2008.



Fonte: Banco de Dados SIGRA

Pelo gráfico acima, observa-se que a disciplina Probabilidade e Estatística apresentou maiores índices de reprovação na maioria dos semestres analisados. Ainda sobre esta disciplina, nota-se que o índice de reprovação cresceu ao longo do período, sendo que mais acentuadamente a partir do segundo semestre de 2005 até o primeiro de 2007, quando começa a decrescer.

A disciplina Bioestatística apresentou no segundo semestre de 2004 o maior índice de reprovação comparado com as demais, seguido de uma queda acentuada até o segundo semestre de 2005.

Percebe-se também um comportamento parecido entre as disciplinas Bioestatística e Estatística Aplicada no período do segundo semestre de 2006 até o segundo de 2008.

A partir da breve análise exposta acima, é importante traçar as características dos estudantes que cursaram as disciplinas Bioestatística, Estatística Aplicada e Probabilidade e Estatística tanto no quesito de reprovação quanto aprovação. Assim sendo, a variável resposta de interesse definida no modelo será reprovação dos alunos, sendo esta considerada de duas formas:

- Reprovação geral – alunos que tiveram menções iguais a MI, II e SR e;
- Reprovação presencial – alunos com menções iguais a MI e II (desconsiderando abandonos).

As tabelas a seguir apresentam as medidas de associação entre reprovação sob os dois focos e diversas variáveis, as quais inspiram possibilidade de dependência com a variável resposta. Variáveis como “frequência do aluno”, “Teve mais de um professor”, “créditos das disciplinas BIO, EA e PE” e “forma de saída da UnB” não foram investigadas por não possuírem fidedignidade com a realidade (não são confiáveis), por não contribuírem para o enriquecimento do estudo ou por fazerem alusão a situações futuras.

Tabela 5 – Análise Bivariada entre a reprovação geral e possíveis variáveis explicativas (continua).

Variáveis	% (n=10560)	% reprovados	Valor da Estatística	Valor-p
Sexo				
Feminino	35,01	20,21	150,0098	<0,0001
Masculino	64,99	31,36		
Cidades				
Centro	66,38	26,69		
Cidades Consolidadas	20,47	28,4		
Assentamentos Recentes	10,09	29,58	12,407	0,0146
Entorno Imediato	1,17	37,9		
Outras	1,88	26,13		
Brasileiro				
Sim	97,45	27,05	32,4368	<0,0001
Não	2,55	42,75		
Semestre de Ingresso				
1	50,71	28,96	12,4534	0,0004
2	49,29	25,9		
Forma de Ingresso				
PAS	21,74	21,39		
Vestibular	70,69	27,82		
Acordos/Convênio/Cortesia	1,94	50,24	133,8816	<0,0001
Mobilidade	1,18	33,6		
Transferências	4,44	39,66		
Período e Ano de Ingresso				
Anterior a 2001	4,73	40,48		
2001	3,72	38,93		
2002	5,72	34,77		
2003	13,94	29,48		
2004	17,59	25,3	159,4865	<0,0001
2005	18,1	28,52		
2006	16,57	27,26		
2007	13,61	20,88		
2008	6,02	16,98		
Tipo do curso				
Licenciatura	6,79	39,61		
Bacharelado	57,13	26,31	57,6172	<0,0001
Profissional	36,08	26,98		
Mudança de opção				
Sim	0,81	29,07	0,1138	**0,7358**
Não	99,19	27,44		

Tabela 5 – Análise Bivariada entre a reprovação geral e possíveis variáveis explicativas (Conclusão).

Variáveis	% (n=10560)	% reprovados	Valor da Estatística	Valor-p
EA	53,58	25,86		
PE	31,16	32,85	79,0512	<0,0001
Bio	15,26	22,04		
Modalidade				
Obrigatória	90,88	27,58		
Optativa	8,05	22,82	41,5665	<0,0001
Módulo livre	1,07	51,33		
Fez disciplina no fluxo				
Sim	45,11	19,1		
Não	44,27	36,81	379,5528	<0,0001
Desconhecido	10,62	23,91		
Turno da disciplina				
Matutino	44,8	24,56		
Vespertino	26,67	25,53	85,7691	<0,0001
Noturno	28,53	33,79		
Quantas vezes cursou a disciplina				
1 vez	82,97	24,63		
2 vezes	12,77	42,18	208,8662	<0,0001
3 vezes	3,34	37,68		
4 vezes ou mais	0,91	40,63		
Período da disciplina				
20041	10,19	23,23		
20042	10,08	21,24		
20051	9,99	27,96		
20052	10,45	25,57		
20061	10,68	32,36	145,4672	<0,0001
20062	10,11	37,17		
20071	10,31	34,07		
20072	9,45	23,05		
20081	9,54	26,81		
20082	9,2	21,91		
Professores	-	-	750,3387	<0,0001
Professor do quadro				
Sim	63,87	25,37	40,7807	<0,0001
Não	36,13	31,14		

Fonte: SIGRA

Tabela 6 – Análise Bivariada entre a reprovação presencial e possíveis variáveis explicativas (continua).

Variáveis	% (n=9172)	% reprovados	Valor da Estatística	Valor-p
Sexo				
Feminino	36,56	12,02	76,2282	<0,0001
Masculino	63,44	19,04		
Cidades				
Centro	66,65	15,93		
Cidades Consolidadas	20,17	16,32		
Assentamentos Recentes	10,1	19,01	13,6875	0,0084
Entorno Imediato	1,14	26,67		
Outras	1,94	17,42		
Brasileiro				
Sim	97,53	16,08	41,6132	<0,0001
Não	2,47	32,16		
Semestre de Ingresso				
1	50,39	17,7	10,1429	0,0014
2	49,61	15,23		
Forma de Ingresso				
PAS	22,55	12,72		
Vestibular	70,45	16,62		
Acordos/Convênio/Cortesia	1,84	39,64	100,3235	<0,0001
Mobilidade	1,16	21,7		
Transferências	4	22,89		
Período e Ano de Ingresso				
Anterior a 2001	4,24	23,65		
2001	3,23	18,92		
2002	5,3	18,93		
2003	13,64	17,03		
2004	17,56	13,84	47,8483	<0,0001
2005	17,95	17,01		
2006	17,07	18,71		
2007	14,53	14,7		
2008	6,48	11,11		
Tipo do curso				
Licenciatura	6,18	23,63		
Bacharelado	57,04	15,02	31,8153	<0,0001
Profissional	36,77	17,52		
Mudança de opção				
Sim	0,82	18,67	0,2642	**0,6073**
Não	99,18	16,46		

Tabela 6 – Análise Bivariada entre a reprovação presencial e possíveis variáveis explicativas (conclusão).

Variáveis	% (n=9172)	% reprovados	Valor da Estatística	Valor-p
Disciplina				
EA	53,33	14,23		
PE	30,85	21,91	88,3413	<0,0001
Bio	15,82	13,44		
Modalidade				
Obrigatória	91,22	16,94		
Optativa	8,02	10,87	18,8261	<0,0001
Módulo livre	0,75	20,29		
Fez disciplina no fluxo				
Sim	48,38	13,14		
Não	41,2	21,83	137,2436	<0,0001
Desconhecido	10,42	10,77		
Turno da disciplina				
Matutino	45,88	15,19		
Vespertino	27,58	17,11	10,1495	0,0063
Noturno	26,54	18,04		
Quantas vezes cursou a disciplina				
1 vez	84,69	14,98		
2 vezes	11,39	25,36	83,3784	<0,0001
3 vezes	3,09	22,26		
4 vezes ou mais	0,83	25		
Período da disciplina				
20041	10,27	12,31		
20042	10,6	13,79		
20051	9,66	14,22		
20052	10,38	13,76		
20061	10,19	18,4	82,5119	<0,0001
20062	8,96	18,37		
20071	10,4	24,74		
20072	9,98	16,07		
20081	9,87	18,56		
20082	9,69	14,62		
Professores	-	-	565,6809	<0,0001
Professor do quadro				
Sim	65,34	16	2,8001	0,0943
Não	34,66	17,36		

Fonte: SIGRA

Observando as análises bivariadas acima, percebem-se que a maioria das variáveis apresentou dependência significativa com os dois tipos de reprovação considerados, logo estas serão candidatas aos modelos logísticos que serão gerados posteriormente. Ao observar o p-valor da variável “Mudança de opção” nas duas tabelas (fonte em negrito), nota-se que este foi bastante elevado, aceitando a hipótese nula de independência.

Pela segunda tabela de associações é possível ver que a variável “Professor do quadro” também não será descartada de imediato, já que a significância comumente utilizada nesta fase nos modelos logísticos ($\alpha = 25\%$) é maior que a dos demais modelos ($\alpha = 5\%$).

6.3- Modelagem

Primeiramente, vale ressaltar que na elaboração de todos os modelos logísticos foi utilizado o método de seleção *Stepwise*, com 5% de significância para entrada e saída de variáveis.

Ao gerar os modelos logísticos, verificou-se que estes foram compostos, em sua maioria, pelas mesmas variáveis explicativas, como: “Sexo”, “Brasileiro” (sim ou não), “Fez disciplina no fluxo”, “Modalidade”(obrigatória e etc), “Ordem que cursou”, “Curso”, “Período de Ingresso”, “Professor” e “Período que fez a disciplina”. Além disto, ambos os modelos rejeitaram a inclusão das seguintes variáveis: “Cidade”, “Semestre de ingresso na UnB” e “Tipo de Curso”.

Inicialmente, a variável “Tipo de Curso” foi categorizada em “Licenciatura”, “Bacharelado” e cunho “Profissional”, porém verificou-se que o padrão do Bacharelado aproximava-se do Profissional. Assim sendo, tomou-se a iniciativa de categorizá-las de novo em Licenciatura ou não. Mesmo após tal procedimento, esta variável não foi incluída nos modelos.

Quanto às divergências entre os modelos, destacam-se as variáveis “Disciplina”, “Turno” e “Professor do Quadro”, sendo que as duas primeiras são referentes ao modelo de reprovação geral.

Tabela 7: Resumo da Seleção Stepwise para os modelos de Reprovação Geral e Presencial

Variáveis	Geral	Presencial
Sexo	X	X
Brasileiro	X	X
Curso	X	X
Período de ingresso na UnB	X	X
Disciplina	X	
Período que cursou a disciplina	X	X
Ordem que cursou a disciplina	X	X
Turno	X	
Modalidade	X	X
Fez disciplina no fluxo	X	X
Professor	X	X
Professor do quadro		X
Número de SR's		X

Para a elaboração dos modelos, testou-se a hipótese nula do Teste Razão de Verossimilhança, no intuito de verificar se existe regressão. As hipóteses, como visto anteriormente, foram:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \\ H_a: pelo\ menos\ um\ \beta_k \neq 0\ de\ H_0 \end{cases}$$

O p-valor ($\hat{\alpha} < 0{,}0001$) referente à estatística Qui-quadrado revelou que existe pelo menos um $\beta_k \neq 0$, ou seja, existe regressão para ambos os modelos.

Os modelos serão compostos pela seguinte estrutura: $\hat{\pi}_{i'} = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1}$. As tabelas com os coeficientes $b_k$estimados de cada categoria das variáveis selecionadas pelos testes de Wald estão no apêndice. Vale frisar que $\hat{\pi}_{i'}$ é a transformação logito, logo para se saber a resposta média, deve-se isolar $\hat{\pi}_i$ em $\hat{\pi}'_i$, assim $E(Y)$ será:

$$\hat{\pi}_i = \frac{\exp\left(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1}\right)}{1 + \exp\left(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1}\right)}$$

Tabela 8: Teste de Adequabilidade de Ajustamento de Hosmer-Lemeshow para os modelos.

Tipo de Reprovação	Estatística Qui-quadrado	Graus de liberdade	Pr> Qui-quadrado
Geral	11,9354	8	0,1541
Presencial	7,9534	8	0,4380

O teste de Adequabilidade do Modelo de Hosmer-Lemeshow indicou pela não rejeição da hipótese nula, para ambos os modelos, a um nível de 5% de significância, implicando que os modelos foram bem ajustados aos dados.

6.3.1- Modelos por disciplina

Devido ao interesse em se estudar a reprovação em cada disciplina separadamente, o banco foi dividido em três partes, gerando-se um modelo para cada disciplina. Este modelo abrange melhor a individualidade de cada disciplina, visto que os cursos não são os mesmos, mudando assim as características dos alunos. Vale frisar que estes três modelos só terão foco na reprovação presencial.

Ao gerar os modelos de reprovação presencial para cada disciplina, constatou-se que estes tiveram as seguintes variáveis em comum: “Sexo”, “Curso”, “Período de Ingresso”, “Professor” e “Período”. Já as ausentes em todos eles foram “Cidade” e “Licenciatura”.

Tabela 9 : Resumo da Seleção Stepwise para Reprovação Presencial nas três disciplinas

Variáveis	PE	BIO	EA
Sexo	X	X	X
Brasileiro			X
Curso	X	X	X
Período de ingresso na UnB	X	X	X
Período que cursou a disciplina	X	X	X
Ordem que cursou a disciplina	X		X
Turno			X
Modalidade			X
Fez disciplina no fluxo	X	X	
Professor	X	X	X
Professor do quadro			X
Número de SR's			X

Tabela 10: Resumo do Teste de Razão de Verossimilhança.

Disciplinas	Estatística Qui-quadrado	Graus de liberdade	Pr> Qui-quadrado
BIO	240,7758	32	<0.0001
PE	474,8545	47	<0.0001
EA	523,7951	64	<0.0001

Ainda sobre a hipótese nula ($H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$), o p-valor da estatística Qui-quadrado do Teste de Razão de Verossimilhança informou que há evidências para dizer que pelo menos um β_k é diferente de zero, ou seja, existe regressão.

No que tange a adequabilidade dos modelos, o Teste de Hosmer-Lemeshow revelou que todos os três modelos testados se ajustaram bem aos dados, isto é verificado nos p-valores bastante elevados, revelando assim que há evidências para se dizer que os modelos se ajustaram bem aos dados.

Tabela 11: Teste de Adequabilidade de Ajustamento de Hosmer-Lemeshow para os três modelos

Disciplinas	Estatística Qui-quadrado	Graus de liberdade	Pr> Qui-quadrado
BIO	3,9775	8	0,8592
PE	4,3706	8	0,8222
EA	3,9622	8	0,8650

Todos os modelos estudados terão a seguinte estrutura: $\hat{\pi}_{i'} = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1}$. Sendo que $\hat{\pi}_{i'}$ é a transformação *logito*, ou seja:

$$\hat{\pi}'_i(x_j) = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1}$$

Assim, os modelos logitos serão compostos pelos betas estimados (b_k) nas tabelas de cada respectivo modelo no apêndice provenientes dos testes de Wald. A exemplificar, o modelo logístico para a reprovação presencial da disciplina Bioestatística será:

Tabela 12: Variáveis do modelo de reprovação presencial de Bioestatística

Variável	Categoria de Referência	Correspondente X	j
Fez no Fluxo	Sim	X1j	1 - não e 2 - outro
Período	1º de 2004	X2j	1 - 2º/2004 , ... , 9 -2º/2008
Professor	35º	X3j	1 - 5º , 2 -11º, 3-14º, 4-15º, 5-17º, 6-20º, 7-22º, 8 -29º, 9 -32º
Sexo	Masculino	X4j	1 - Feminino
Curso	Outros	X5j	1 - Agronomia, 2- Engenharia Florestal, 3- Medicina Veterinária
Período de Ingresso	Anterior a 2001	X6j	1 - 2001 , ... , 8 - 2008

$$\hat{\pi}'_i(x_j) = -2{,}2391 + 0{,}5404X_{11} + 0{,}1552X_{12} + 0{,}3612X_{21} - 1{,}7638X_{22} - 1{,}3615X_{23} - 0{,}6163X_{24} - 0{,}192X_{25} + 0{,}1432X_{26} - 0{,}1653X_{27} + 1{,}2096X_{28} + 1{,}6845X_{29} + 0{,}5847X_{31} + 0{,}1878X_{32} + 1{,}6674X_{33} - 0{,}00907X_{34} + 0{,}8802X_{35} - 3{,}871X_{36} - 0{,}4414X_{37} - 0{,}4436X_{38} - 0{,}5343X_{39} - 0{,}3561X_{41} + 0{,}05X_{51} + 0{,}7875X_{52} + 0{,}1673X_{53} + 0{,}9517X_{61} + 0{,}6915X_{62} + 6391X_{63} + 0{,}3843X_{64} - 0{,}25X_{65} - 0{,}8357X_{66} - 1{,}1291X_{67} - 1{,}0904X_{68}$$

Para se saber a resposta média, deve-se transformar $\hat{\pi}'_i(x_j)$, isolando $\hat{\pi}_i$, assim $E(Y)$ será:

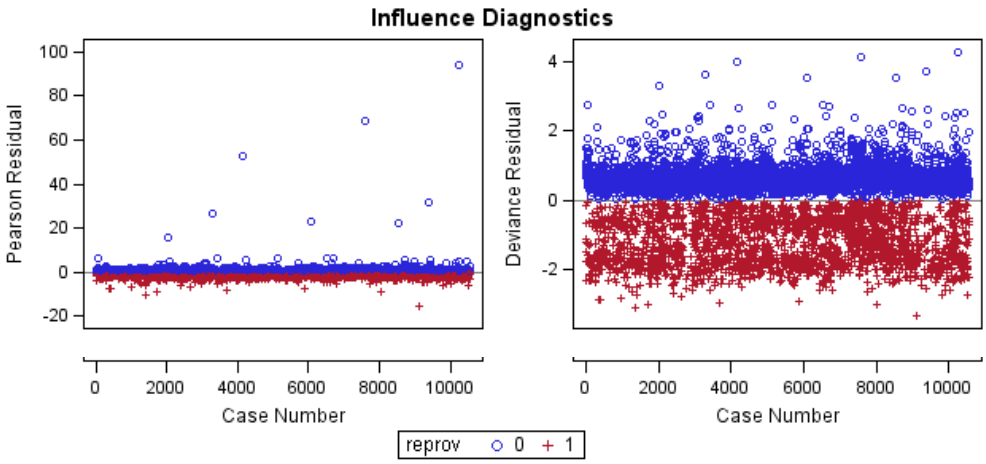
$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1})}{1 + \exp(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \cdots + b_{p-1} x_{i,p-1})}$$

6.4- Análises de Resíduo

6.4.1- Reprovação Geral

A presente análise de resíduos se refere ao modelo geral de Reprovação com menção SR. O Gráfico de influências por dados observados, mostra que há certa homogeneidade na reprovação dos alunos, e uma singela heterogeneidade entre os aprovados, visto que a nuvem de pontos desta ultima esta menos condensada. Nota-se, também, que há poucos resíduos *deviance* entre os aprovados que foram maiores do que dois, revelando assim poucas observações discrepantes. Já entre os reprovados, percebe-se que muitos resíduos Deviance foram superiores a 2, porém estes não aparentam ser valores discrepantes visto que não estão isolados.

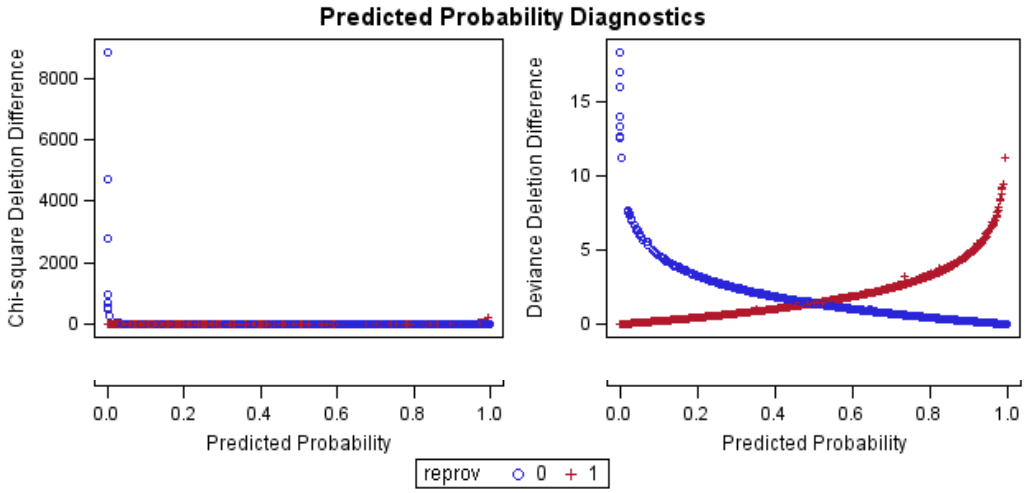
Gráfico 10: Resíduos de Pearson e Deviance por observações para o modelo com Reprovação Geral



O gráfico de Resíduos de Pearson ressalta visualmente melhor a possível presença de valores discrepantes entre aqueles que foram aprovados.

O gráfico abaixo plota as probabilidades estimadas versus a diferença entre o Qui-quadrado de adequabilidade de ajuste. Através deste, também é possível investigar pontos influentes no modelo.

Gráfico 11: Diferença entre o Qui-quadrado versus Probabilidades estimadas para o modelo de reprovação Geral.



Nota-se nestes dois gráficos acima que quando a probabilidade predita é baixa, há cerca de três pontos discrepantes entre os que foram aprovados. À medida que a probabilidade predita cresce, o comportamento entre os reprovados e aprovados parecem ser semelhantes. Retirando as possíveis observações discrepantes, nota-se que as inclinações são bem semelhantes entre os reprovados e aprovados.

6.4.2- Reprovação Presencial

6.4.2.1 – Estatística Aplicada, Bioestatística e Probabilidade e Estatística.

No que tange a reprovação presencial, o Gráfico de influências por dados observados, mostra que há certa homogeneidade na aprovação dos alunos, visto que a nuvem de pontos está bem condensada. Dentre as três disciplinas, a que apresentou maior dispersão foi a Bioestatística, nota-se também que a quantidade de observações nesta disciplina é bem menor se comparada com as outras.

Nota-se, também, que há poucos resíduos *deviance* que foram maiores do que 2 em P.E., tal fato não ocorre na disciplina E.A., pois nota-se que a distribuição dos pontos parece estar centrada no resíduo deviance igual a -2, ou seja, cerca de 50% destas observações aparentam ser discrepantes,

podendo ter uma influência no modelo, porém como estas não estão isoladas, estas observações não são discrepantes.

Quanto à nuvem de pontos, nota-se que esta é mais heterogênea entre os aprovados do que entre os reprovados. É interessante salientar que tal fato é contrário ao modelo geral com reprovação incluindo os SR's, visto que este último apresenta valores mais atípicos entre os aprovados. Esta diferença entre os modelos se dá na particularidade entre as disciplinas e no abandono (SR). Conclui-se então que as observações que aparentavam ser discrepantes no modelo geral são omitidas a desconsiderar os abandonos e as diferenças entre as disciplinas.

Gráfico 12: Resíduos de Pearson e Deviance por observações para o modelo com Reprovação Presencial da disciplina **E.A.**

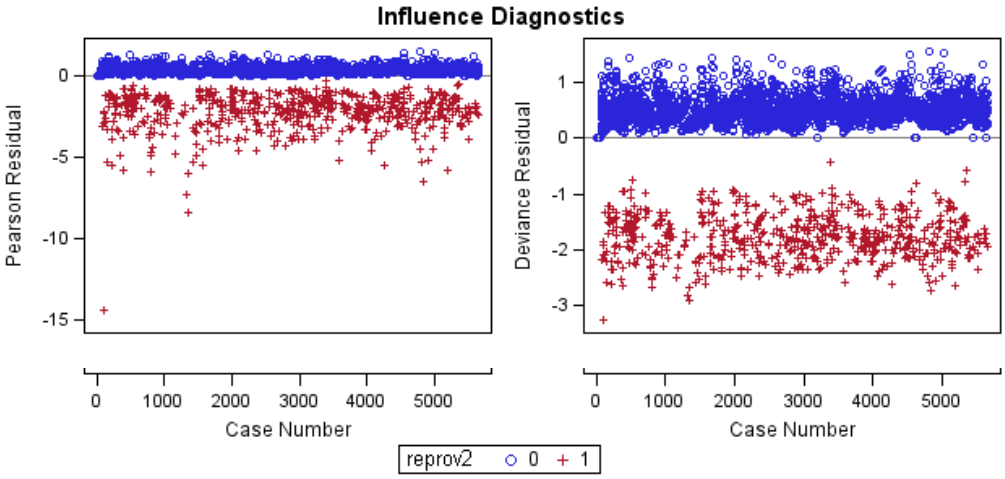


Gráfico 13: Resíduos de Pearson e Deviance por observações para o modelo com Reprovação Presencial da disciplina **BIO.**

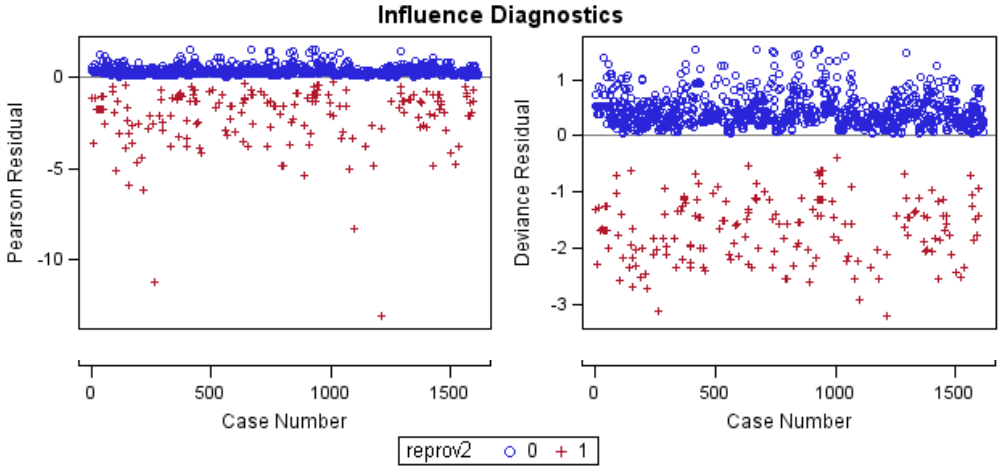
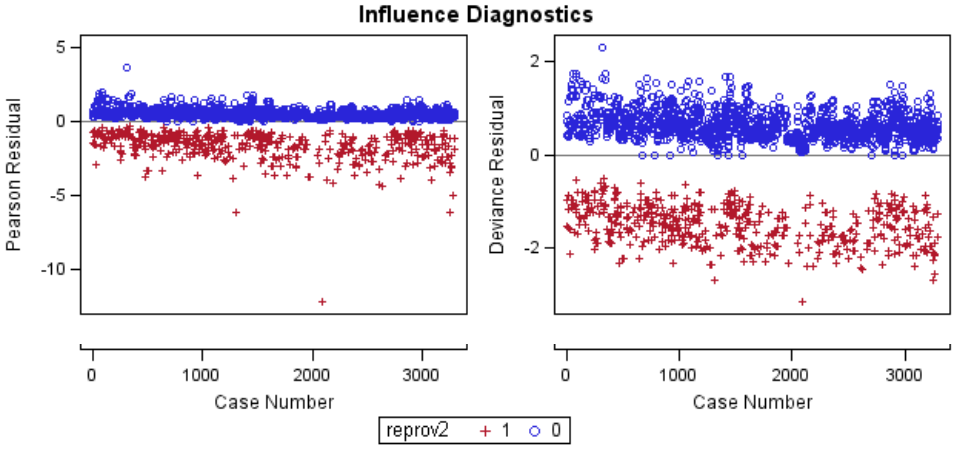


Gráfico 14: Resíduos de Pearson e Deviance por observações para o modelo com Reprovação Presencial da disciplina **P.E.**



O gráfico de Resíduos de Pearson ressalta visualmente melhor a diferença da homogeneidade a forte homogeneidade entre os aprovados. Este gráfico aponta uma possível observação discrepante entre os reprovados.

Gráfico 15: Diferença entre o Qui-quadrado versus Probabilidades estimadas para **E.A.**

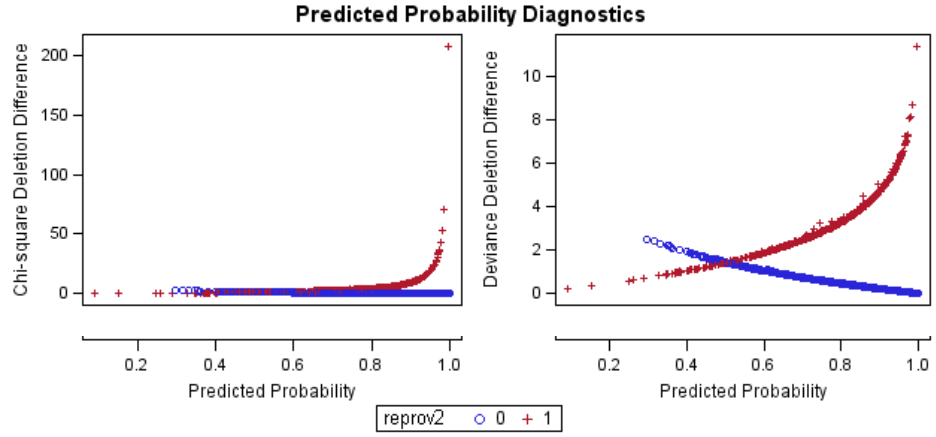


Gráfico 16: Diferença entre o Qui-quadrado versus Probabilidades estimadas para **BIO.**

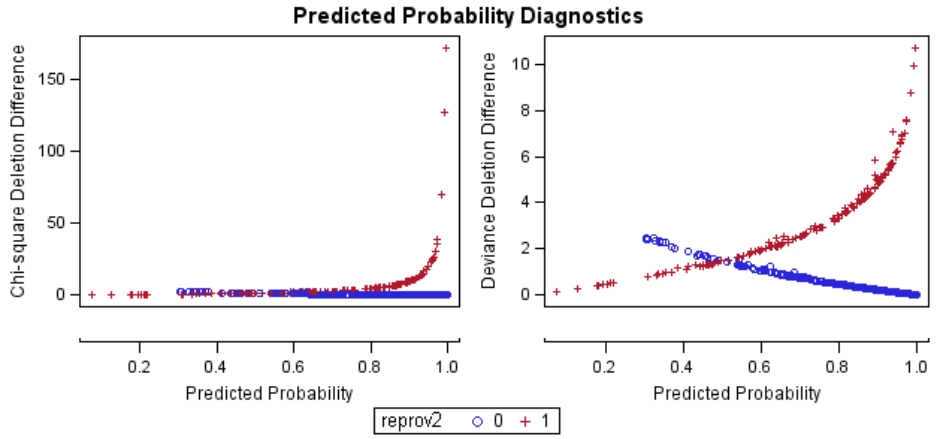
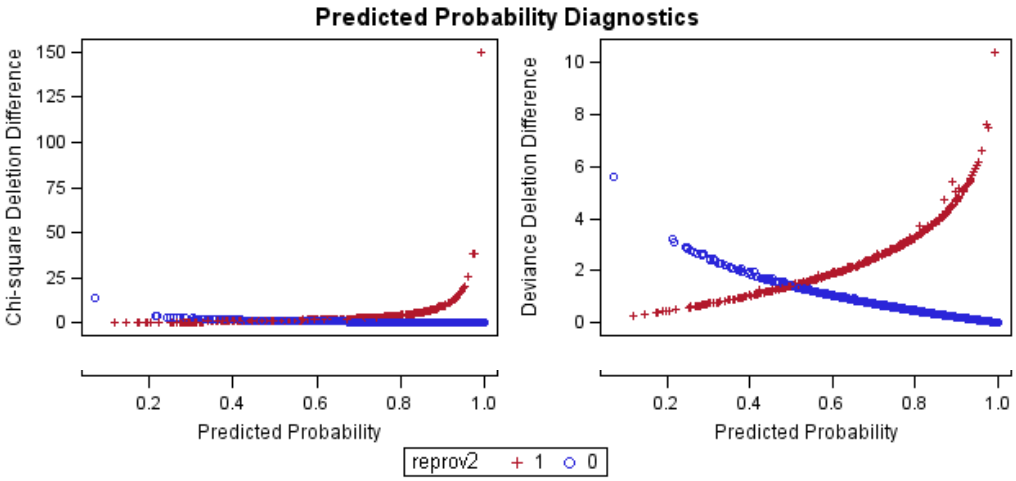


Gráfico 17: Diferença entre o Qui-quadrado versus Probabilidades estimadas para **P.E.**



O gráfico acima plota as probabilidades estimadas versus a diferença entre o Qui-quadrado de adequabilidade de ajuste. Através deste, é possível investigar pontos influentes no modelo. Novamente, notam-se dois comportamentos distintos entre os aprovados e reprovados à medida que a probabilidade predita cresce. Percebe-se uma inclinação aguçada nos pontos referentes aos reprovados, indicando duas possíveis observações discrepantes à medida que a probabilidade estimada cresce.

6.5- Odds Ratio ou Razão de Chances

Tabela 13 – *Odds ratio* do modelo de reprovação presencial da disciplina Bioestatística

Odds Ratio Estimates				
Effect		*Point Estimate*	*95% Wald Confidence Limits*	
SEXO	*F vs M*	0.491	0.337	0.713
fez_no_fluxo	*nao vs sim*	3.442	2.184	5.425
fez_no_fluxo	*outro vs sim*	2.342	0.767	7.149
curso2	*Agronomia vs Outros*	2.871	1.021	8.073
curso2	*Engenharia Florestal vs Outros*	6.004	2.094	17.217
curso2	*Medicina Veterinária vs Outros*	3.229	1.112	9.377
dataingressocateg 2001 vs anterior a 2001		1.356	0.426	4.314

Tabela 13 – *Odds ratio* do modelo de reprovação presencial da disciplina Bioestatística (continuação).

Odds Ratio Estimates				
Effect		*Point Estimate*	*95% Wald Confidence Limits*	
dataingressocateg 2002 vs anterior a 2001		1.045	0.403	2.711
dataingressocateg 2003 vs anterior a 2001		0.992	0.418	2.352
dataingressocateg 2004 vs anterior a 2001		0.769	0.300	1.970
dataingressocateg 2005 vs anterior a 2001		0.404	0.145	1.125
dataingressocateg 2006 vs anterior a 2001		0.227	0.074	0.694
dataingressocateg 2007 vs anterior a 2001		0.169	0.046	0.629
dataingressocateg 2008 vs anterior a 2001		0.176	0.035	0.891
Nome_Professor	*5 vs 35*	0.248	0.053	1.168
Nome_Professor	*11 vs 35*	0.167	0.041	0.681
Nome_Professor	*14 vs 35*	0.732	0.187	2.875
Nome_Professor	*15 vs 35*	0.137	0.044	0.422
Nome_Professor	*17 vs 35*	0.333	0.114	0.978
Nome_Professor	*20 vs 35*	0.003	<0.001	0.031
Nome_Professor	*22 vs 35*	0.089	0.018	0.450
Nome_Professor	*29 vs 35*	0.089	0.024	0.331
Nome_Professor	*32 vs 35*	0.081	0.019	0.339
PERIODO	*20042 vs 20041*	0.712	0.262	1.933
PERIODO	*20051 vs 20041*	0.085	0.026	0.274
PERIODO	*20052 vs 20041*	0.127	0.044	0.365
PERIODO	*20061 vs 20041*	0.268	0.089	0.805
PERIODO	*20062 vs 20041*	0.410	0.137	1.224
PERIODO	*20071 vs 20041*	0.573	0.176	1.861
PERIODO	*20072 vs 20041*	0.421	0.128	1.386
PERIODO	*20081 vs 20041*	1.664	0.498	5.557
PERIODO	*20082 vs 20041*	2.675	0.689	10.389

Fonte:SIGRA

As estimativas pontuais das razões de chances para o modelo de reprovação presencial da disciplina Bioestatística mostraram que a chance das mulheres reprovar é menor que a chance dos homens e que a chance de reprovar dos estudantes que não fizerem a disciplina no fluxo é 3,442 vezes a chance de reprovar dos alunos que fizeram a disciplina no fluxo.

Quanto aos professores, nenhum professor apresentou maior chance de reprovação comparada com a chance de reprovação do professor 35. Vale lembrar que os professores foram intitulados com números de 1 a 43 para o sigilo de suas identidades.

Quanto ao ano de ingresso na UnB, a chance de reprovar de um estudante que entrou na UnB em 2008 é 0,176 da chance de um estudante que entrou antes de 2001 e a chance de reprovação de um estudante que ingressou na UnB em 2001 é 35,6% maior que a chance de um que ingressou em algum ano anterior a 2001.

O curso que apresentou maior chance de reprovar comparado com a chance de reprovar do curso de Agronomia foi Engenharia Florestal seguido pelo curso de Medicina Veterinária. A chance de reprovação da categoria “Outros cursos” é 0,348 da chance de reprovação dos estudantes que pertencem ao curso de Agronomia.

Para os demais modelos gerados, podem-se fazer tais afirmativas de forma análoga a partir das tabelas no apêndice. Vale ressaltar a importância de observar os intervalos de confiança das razões de chances estimadas: a presença do valor 1 dentro do intervalo requer mais cuidado já que este representa igualdade de chances entre as categorias.

7 - Conclusões

Os resultados iniciais da pesquisa mostraram que a disciplina Probabilidade e Estatística obteve o maior índice de reprovação dentre os alunos que concluíram as disciplinas (21,91%), seguida por Estatística Aplicada, que apresentou pouca diferença do índice da disciplina Bioestatística (14,23% e 13,44%, respectivamente). A respeito dos alunos que não concluíram as disciplinas, Bioestatística apresentou maior índice, seguida de Probabilidade e Estatística. Quanto aos trancamentos, à disciplina Bioestatística também apresentou o maior índice e Estatística Aplicada, o menor.

A partir da análise descritiva feita, optou-se por gerar diversos modelos devido às particularidades de cada variável resposta. Assim sendo, as variáveis dependentes analisadas e respectivas variáveis preditoras retidas em seus modelos foram:

- Reprovação geral: “Sexo”, “Disciplina”, “Brasileiro (sim ou não)”, ”Fez disciplina no fluxo”, ”Modalidade”, ”Curso”, ”Ordem que cursou a disciplina”, ”Período de ingresso”, ”Professor”, ”Turno” e ”Período que cursou a disciplina”;
- Reprovação presencial: “Sexo”, “Brasileiro (sim ou não)”, ”Fez disciplina no fluxo”, ”Modalidade”, ”Curso”, ”Ordem que cursou a disciplina”, ”Período de ingresso”, ”Professor”, ”Professor do quadro”, ”Período que cursou a disciplina” e “número de SR’s”;
- Reprovação presencial de Bioestatística: “Sexo”, ”Fez disciplina no fluxo”, ”Curso”, ”Período de ingresso”, ”Professor” e ”Período que cursou a disciplina”;
- Reprovação presencial de Probabilidade e Estatística: “Sexo”, ”Fez disciplina no fluxo”, ”Ordem que cursou a disciplina”, ”Curso”, ”Período de ingresso”, ”Professor” e ”Período que cursou a disciplina” e;
- Reprovação presencial de Estatística Aplicada: “Sexo”, ”Modalidade”, “Brasileiro (sim ou não)”, ”Curso”, “Ordem que cursou a

disciplina", "Período de ingresso", "Professor", "Turno", "Professor do quadro", "Período que cursou a disciplina" e "Números de SR's".

Para todos os modelos citados, os testes de Razão de Verossimilhança indicaram que há evidências de que, realmente, existe regressão. O teste de Hosmer-Lameshow de ajustamento indicou que todos os modelos se ajustaram bem aos dados.

Verificou-se que para todas as três disciplinas, a chance de reprovação das mulheres é menor que a chance de reprovação dos homens, sendo que a disciplina Bioestatística apresentou a menor razão destas $(\theta = 0{,}49)$.

A disciplina Estatística Aplicada foi a única que apresentou a variável "turno" em seu modelo, mostrando que a chance de reprovar dos estudantes que cursam a esta disciplina durante à noite é quase três vezes maior que a chance de reprovar dos que cursam durante a manhã. Em relação ao vespertino, a chance de reprovar 70% maior do que a chance de reprovar no matutino.

Quantos aos professores, a chance de reprovar na disciplina Bioestatística com o professor 35 é maior do que as chances de todos os outros professores que ministraram essa disciplina. Em Estatística Aplicada., a maior razão de chances entre os professores foi de 54,197, sendo esta a comparação entre o professor 34 e o 42(referência). Já em Probabilidade e Estatística., a chance de reprovar com o professor 35 é cerca de treze vezes maior do que a chance de reprovar com o professor de referência (43).

Conclui-se então, que os alunos que cursaram as três disciplinas oferecidas pelo Departamento de Estatística, reprovaram segundo algumas características comuns: "Sexo, "Professor", "Curso" "Período que cursou a disciplina" e "Período que ingressou na UnB" e algumas específicas para cada disciplina.

Apêndices

Apêndice 1 – Detalhamento dos cursos que fazem Bioestatística, Estatística Aplicada e Probabilidade e Estatística.

Disciplinas	Cursos
Bioestatística	Agronomia
	Engenharia Florestal
	Medicina Veterinária
Estatística Aplicada	Administração Diurno
	Administração Noturno
	Antropologia
	Arquivologia
	Biblioteconomia
	Ciência Política
	Ciências Ambientais
	Ciências Contábeis
	Ciências Sociais
	Geografia
	Psicologia
	Relações Internacionais
	Sociologia
Probabilidade e Estatística	Ciência da Computação
	Computação
	Engenharia Civil
	Engenharia da Computação
	Engenharia de Produção
	Engenharia de Redes
	Engenharia Elétrica
	Engenharia Mecânica
	Engenharia Mecatrônica
	Matemática

Apêndice 2 – Categorização das formas de ingresso na UnB

Categoria	Formas de Ingresso
PAS	Programa de Avaliação Seriada
Vestibular	Vestibular
Mobilidade	Duplo Curso
	Dupla Habilitação
	Mudança de Curso

	Mudança de Habilitação
Transferências	Transferência Obrigatória Transferência Facultativa
Acordos e Convênios	Acordo Cultural-PEC Matrícula Cortesia Convênio-Int Convênio - Andifes

Apêndice 3 – Categorização da variável “cidade onde mora”.

Categoria	**Cidades**
Centro	Brasília Lago Norte Lago Sul / Jardim Botânico Cruzeiro/SMU Sudoeste Octogonal Vila Planalto
Cidades Consolidadas	Gama Guará/Lúcio Costa Taguatinga Candangolândia Sobradinho Núcleo Bandeirante Park Way
Assentamentos Recentes	Brazlândia Ceilândia Paranoá Planaltina Recanto das Emas Riacho Fundo Samambaia

	Santa Maria
	São Sebastião
	Arniqueiras
	Águas Claras
	Vicente Pires
Entorno Imediato	Água Fria de Goiás
	Águas Lindas de Goiás
	Cidade Ocidental
	Luziânia
	Novo Gama
	Padre Bernardo
	Planaltina
	Santo Antônio do Descoberto
	Valparaíso de Goiás
Outras	Demais municípios da RIDE* e do estado de Goiás

* A Região Integrada de Desenvolvimento do Entorno (RIDE) corresponde à área de influência mais direta do DF, com características de região metropolitana. A RIDE é formada pelos municípios de Abadiânia, Água Fria de Goiás, Águas Lindas de Goiás, Alexânia, Cabeceiras, Cidade Ocidental, Cocalzinho de Goiás, Cristalina, Formosa, Luziânia, Mimoso de Goiás, Novo Gama, Padre Bernardo, Pirenópolis, Planaltina de Goiás, Santo Antônio do Descoberto, Valparaíso de Goiás e Vila Boa no Estado de Goiás e de Unaí, Buritis e Cabeceira Grande no Estado de Minas Gerais.

Apêndice 4 – Parâmetros estimados para o modelo completo tendo como variável resposta a reprovação geral:

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.9168	0.2944	9.6990	0.0018
SEXO	**F**	1	-0.2621	0.0282	86.2023	<.0001
DISCIPLINA	**BIO**	1	-0.4282	0.1217	12.3781	0.0004
DISCIPLINA	**EA**	1	-0.0813	0.1126	0.5214	0.4703
brasileiro	**nao**	1	0.4579	0.0721	40.2957	<.0001
fez_no_fluxo	**nao**	1	0.5177	0.2968	3.0420	0.0811
fez_no_fluxo	**outro**	1	-0.3488	0.5921	0.3470	0.5558
MODALIDADE	**ML**	1	1.0699	0.1515	49.8977	<.0001
MODALIDADE	**OPT**	1	-0.6013	0.1041	33.3469	<.0001
ordem_q_cursou	**2vez**	1	0.3182	0.0806	15.5894	<.0001
ordem_q_cursou	**3vez**	1	-0.2543	0.1089	5.4505	0.0196

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
ordem_q_cursou	**4+vez**	1	-0.3454	0.1792	3.7151	0.0539
curso2	**Administração Diurno**	1	-0.0458	0.1360	0.1135	0.7362
curso2	**Administração Noturn**	1	0.0968	0.1382	0.4906	0.4837
curso2	**Agronomia**	1	0.2476	0.1804	1.8832	0.1700
curso2	**Antropologia**	1	0.4306	0.1750	6.0535	0.0139
curso2	**Arquivologia**	1	0.3552	0.1331	7.1202	0.0076
curso2	**Biblioteconomia**	1	1.0327	0.1446	50.9761	<.0001
curso2	**Ciência Política**	1	0.0141	0.1574	0.0080	0.9288
curso2	**Ciência da Computaçã**	1	-0.4618	0.1633	7.9956	0.0047
curso2	**Ciências Contábeis**	1	0.00636	0.1257	0.0026	0.9597
curso2	**Ciências Sociais**	1	0.6899	0.1548	19.8612	<.0001
curso2	**Computação**	1	-0.0756	0.1649	0.2099	0.6468
curso2	**Engenharia Civil**	1	-0.3287	0.1487	4.8891	0.0270
curso2	**Engenharia Elétrica**	1	-0.9475	0.1845	26.3738	<.0001
curso2	**Engenharia Florestal**	1	0.3756	0.1785	4.4280	0.0354
curso2	**Engenharia Mecatrôni**	1	-0.8247	0.2005	16.9157	<.0001
curso2	**Engenharia Mecânica**	1	-1.1064	0.1695	42.5826	<.0001
curso2	**Engenharia de Redes**	1	-0.4153	0.1799	5.3309	0.0210
curso2	**Geografia**	1	0.5861	0.1457	16.1709	<.0001
curso2	**Matemática**	1	0.3415	0.1557	4.8107	0.0283
curso2	**Medicina Veterinária**	1	-0.0654	0.2070	0.0998	0.7520
curso2	**Outros**	1	0.3673	0.8538	0.1851	0.6670
curso2	**Relações Internacion**	1	-0.5564	0.1549	12.9087	0.0003
dataingressocateg	**2001**	1	0.5900	0.1151	26.2754	<.0001
dataingressocateg	**2002**	1	0.3924	0.0966	16.4935	<.0001
dataingressocateg	**2003**	1	0.3623	0.0729	24.6695	<.0001
dataingressocateg	**2004**	1	0.0107	0.0662	0.0259	0.8721
dataingressocateg	**2005**	1	-0.1921	0.0646	8.8369	0.0030
dataingressocateg	**2006**	1	-0.3715	0.0755	24.1914	<.0001
dataingressocateg	**2007**	1	-0.5686	0.0965	34.7190	<.0001
dataingressocateg	**2008**	1	-0.9152	0.1336	46.8925	<.0001
Nome_Professor	**1**	1	0.5971	0.1744	11.7192	0.0006
Nome_Professor	**2**	1	0.5883	0.1776	10.9720	0.0009
Nome_Professor	**3**	1	-1.6157	0.5232	9.5352	0.0020
Nome_Professor	**5**	1	0.0533	0.2149	0.0616	0.8040
Nome_Professor	**6**	1	0.5947	0.1621	13.4565	0.0002
Nome_Professor	**7**	1	1.0672	0.1628	42.9864	<.0001
Nome_Professor	**8**	1	0.5088	0.2145	5.6281	0.0177
Nome_Professor	**10**	1	0.3949	0.3248	1.4783	0.2240
Nome_Professor	**11**	1	-0.1852	0.1762	1.1045	0.2933
Nome_Professor	**13**	1	-0.1110	0.1581	0.4924	0.4829
Nome_Professor	**14**	1	0.1572	0.1812	0.7525	0.3857
Nome_Professor	**15**	1	0.5384	0.2016	7.1325	0.0076
Nome_Professor	**17**	1	0.6290	0.1190	27.9152	<.0001
Nome_Professor	**18**	1	-1.2803	0.3263	15.3919	<.0001
Nome_Professor	**20**	1	-0.5274	0.3298	2.5578	0.1097
Nome_Professor	**22**	1	1.3200	0.2022	42.6049	<.0001
Nome_Professor	**23**	1	-0.2741	0.1541	3.1625	0.0753

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Nome_Professor	**25**	1	0.3722	0.1003	13.7854	0.0002
Nome_Professor	**26**	1	0.0325	0.2773	0.0137	0.9067
Nome_Professor	**27**	1	0.2189	0.1471	2.2136	0.1368
Nome_Professor	**28**	1	-0.3510	0.1737	4.0814	0.0434
Nome_Professor	**29**	1	0.2272	0.1154	3.8807	0.0488
Nome_Professor	**30**	1	-1.4111	0.2078	46.0991	<.0001
Nome_Professor	**31**	1	-0.4214	0.2492	2.8600	0.0908
Nome_Professor	**32**	1	-0.4675	0.3525	1.7593	0.1847
Nome_Professor	**33**	1	-0.1933	0.1620	1.4246	0.2326
Nome_Professor	**34**	1	0.6875	0.1338	26.3901	<.0001
Nome_Professor	**35**	1	1.8685	0.1312	202.6870	<.0001
Nome_Professor	**36**	1	0.6629	0.1807	13.4524	0.0002
Nome_Professor	**37**	1	-0.2930	0.1497	3.8294	0.0504
Nome_Professor	**38**	1	0.0546	0.3493	0.0244	0.8758
Nome_Professor	**39**	1	-0.3998	0.1681	5.6551	0.0174
Nome_Professor	**40**	1	-0.0143	0.1807	0.0062	0.9370
Nome_Professor	**41**	1	-1.1397	0.2392	22.6939	<.0001
Nome_Professor	**42**	1	-2.0701	0.4401	22.1273	<.0001
Turno	**noturno**	1	0.1913	0.0654	8.5594	0.0034
Turno	**vespertino**	1	-0.2873	0.0605	22.5766	<.0001
PERIODO	**20042**	1	-0.2894	0.1040	7.7383	0.0054
PERIODO	**20051**	1	-0.6634	0.0951	48.6463	<.0001
PERIODO	**20052**	1	-0.5851	0.0869	45.3695	<.0001
PERIODO	**20061**	1	-0.0602	0.0826	0.5300	0.4666
PERIODO	**20062**	1	0.2427	0.0793	9.3640	0.0022
PERIODO	**20071**	1	0.4143	0.0878	22.2818	<.0001
PERIODO	**20072**	1	0.0240	0.1021	0.0551	0.8144
PERIODO	**20081**	1	0.6122	0.1062	33.2176	<.0001
PERIODO	**20082**	1	0.7417	0.1288	33.1760	<.0001

Apêndice 5 – Parâmetros estimados para o modelo completo tendo como variável resposta a reprovação presencial:

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.2521	0.3903	33.3026	<.0001
SEXO	**F**	1	-0.2441	0.0359	46.2281	<.0001
brasileiro	**nao**	1	0.5713	0.0837	46.5707	<.0001
fez_no_fluxo	**nao**	1	0.4893	0.3835	1.6281	0.2020
fez_no_fluxo	**outro**	1	-0.4415	0.7644	0.3335	0.5636
MODALIDADE	**ML**	1	0.6546	0.2258	8.4015	0.0037
MODALIDADE	**OPT**	1	-0.5247	0.1444	13.2048	0.0003
n_SR		1	0.4417	0.1005	19.3090	<.0001
ordem_q_cursou	**2vez**	1	0.3855	0.1105	12.1637	0.0005
ordem_q_cursou	**3vez**	1	-0.3650	0.1408	6.7245	0.0095
ordem_q_cursou	**4+vez**	1	-0.6359	0.2479	6.5790	0.0103

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
curso2	Administração Diurno	1	0.00806	0.1499	0.0029	0.9571
curso2	Administração Noturn	1	-0.3263	0.1819	3.2162	0.0729
curso2	Agronomia	1	-0.0825	0.1769	0.2175	0.6410
curso2	Antropologia	1	0.1491	0.2144	0.4836	0.4868
curso2	Arquivologia	1	0.5099	0.1505	11.4822	0.0007
curso2	Biblioteconomia	1	0.8629	0.1672	26.6479	<.0001
curso2	Ciência Política	1	-0.0779	0.1831	0.1808	0.6707
curso2	Ciência da Computaçã	1	-0.1918	0.1616	1.4095	0.2351
curso2	Ciências Contábeis	1	-0.3172	0.1464	4.6924	0.0303
curso2	Ciências Sociais	1	0.3871	0.1863	4.3166	0.0377
curso2	Computação	1	0.1463	0.1643	0.7928	0.3732
curso2	Engenharia Civil	1	-0.0658	0.1386	0.2256	0.6348
curso2	Engenharia Elétrica	1	-0.7050	0.1908	13.6549	0.0002
curso2	Engenharia Florestal	1	0.3108	0.1603	3.7591	0.0525
curso2	Engenharia Mecatrôni	1	-0.4459	0.2039	4.7841	0.0287
curso2	Engenharia Mecânica	1	-0.9381	0.1790	27.4548	<.0001
curso2	Engenharia de Redes	1	-0.0687	0.1785	0.1481	0.7004
curso2	Geografia	1	0.3679	0.1684	4.7716	0.0289
curso2	Matemática	1	0.4529	0.1583	8.1866	0.0042
curso2	Medicina Veterinária	1	-0.1790	0.1989	0.8093	0.3683
curso2	Outros	1	0.2866	1.1032	0.0675	0.7950
curso2	Relações Internacion	1	-0.3929	0.1689	5.4115	0.0200
dataingressocateg	2001	1	0.3576	0.1596	5.0210	0.0250
dataingressocateg	2002	1	0.2922	0.1277	5.2357	0.0221
dataingressocateg	2003	1	0.3735	0.0957	15.2140	<.0001
dataingressocateg	2004	1	-0.0147	0.0869	0.0286	0.8658
dataingressocateg	2005	1	-0.1207	0.0831	2.1065	0.1467
dataingressocateg	2006	1	-0.2580	0.0933	7.6546	0.0057
dataingressocateg	2007	1	-0.4976	0.1189	17.5165	<.0001
dataingressocateg	2008	1	-0.8123	0.1659	23.9706	<.0001
Nome_Professor	1	1	0.9933	0.2039	23.7393	<.0001
Nome_Professor	2	1	1.3162	0.2350	31.3582	<.0001
Nome_Professor	3	1	-1.5590	0.7237	4.6402	0.0312
Nome_Professor	5	1	0.2185	0.2511	0.7573	0.3842
Nome_Professor	6	1	0.2480	0.2277	1.1862	0.2761
Nome_Professor	7	1	1.5169	0.1916	62.6835	<.0001
Nome_Professor	8	1	0.6151	0.2677	5.2791	0.0216
Nome_Professor	10	1	-0.3513	0.4358	0.6496	0.4202
Nome_Professor	11	1	-0.4986	0.2594	3.6941	0.0546
Nome_Professor	13	1	0.1236	0.2072	0.3560	0.5507
Nome_Professor	14	1	0.2317	0.2273	1.0396	0.3079
Nome_Professor	15	1	-0.0693	0.2562	0.0731	0.7869
Nome_Professor	17	1	0.4234	0.1535	7.6132	0.0058
Nome_Professor	18	1	-3.0088	0.9936	9.1706	0.0025
Nome_Professor	20	1	-2.8520	0.9919	8.2675	0.0040
Nome_Professor	22	1	1.4072	0.2398	34.4335	<.0001
Nome_Professor	23	1	-0.3728	0.2127	3.0727	0.0796

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Nome_Professor	25	1	0.2319	0.1449	2.5630	0.1094
Nome_Professor	26	1	0.0847	0.3461	0.0599	0.8067
Nome_Professor	27	1	-0.2293	0.2021	1.2882	0.2564
Nome_Professor	28	1	0.6582	0.2133	9.5236	0.0020
Nome_Professor	29	1	0.1946	0.1491	1.7039	0.1918
Nome_Professor	30	1	-1.1196	0.2491	20.2021	<.0001
Nome_Professor	31	1	-0.3699	0.3212	1.3262	0.2495
Nome_Professor	32	1	-0.6152	0.4369	1.9831	0.1591
Nome_Professor	33	1	-0.7102	0.3251	4.7728	0.0289
Nome_Professor	34	1	1.3416	0.2094	41.0671	<.0001
Nome_Professor	35	1	2.4083	0.2183	121.6542	<.0001
Nome_Professor	36	1	1.2500	0.2564	23.7752	<.0001
Nome_Professor	37	1	-0.3230	0.2367	1.8617	0.1724
Nome_Professor	38	1	0.3470	0.5018	0.4782	0.4892
Nome_Professor	39	1	0.0471	0.2332	0.0408	0.8399
Nome_Professor	40	1	0.3557	0.2628	1.8321	0.1759
Nome_Professor	41	1	-0.1337	0.3161	0.1790	0.6722
Nome_Professor	42	1	-2.3731	0.7283	10.6174	0.0011
professor_quadro	não	1	-0.2741	0.1084	6.4002	0.0114
PERIODO	20042	1	-0.1517	0.1309	1.3428	0.2465
PERIODO	20051	1	-0.8106	0.1285	39.7742	<.0001
PERIODO	20052	1	-0.6168	0.1120	30.3093	<.0001
PERIODO	20061	1	-0.0673	0.1074	0.3923	0.5311
PERIODO	20062	1	-0.0582	0.1061	0.3015	0.5829
PERIODO	20071	1	0.5267	0.1073	24.1134	<.0001
PERIODO	20072	1	0.1386	0.1236	1.2567	0.2623
PERIODO	20081	1	0.6489	0.1284	25.5246	<.0001
PERIODO	20082	1	0.9842	0.1549	40.3614	<.0001

Apêndice 6 – Parâmetros estimados para o modelo da disciplina **E.A.** tendo como variável resposta a reprovação presencial:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3258	28.8257	0.0065	0.9357
SEXO	F	1	-0.2103	0.0466	20.3820	<.0001
brasileiro	nao	1	0.7592	0.0938	65.5374	<.0001
MODALIDADE	ML	1	1.3434	0.3117	18.5749	<.0001
MODALIDADE	OPT	1	-0.8330	0.2853	8.5241	0.0035
n_SR		1	0.4158	0.1461	8.1035	0.0044
ordem_q_cursou	2vez	1	0.6827	0.1593	18.3761	<.0001
ordem_q_cursou	3vez	1	-0.6724	0.2088	10.3718	0.0013
ordem_q_cursou	4+vez	1	-0.4690	0.3441	1.8579	0.1729
curso2	Administração Diurno	1	-0.1321	0.1526	0.7490	0.3868
curso2	Administração Noturn	1	-0.4106	0.1903	4.6539	0.0310
curso2	Agronomia	1	-0.2285	0.7864	0.0844	0.7714

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
curso2	Antropologia	1	0.2467	0.2128	1.3436	0.2464
curso2	Arquivologia	1	0.4312	0.1583	7.4221	0.0064
curso2	Biblioteconomia	1	0.8653	0.1729	25.0436	<.0001
curso2	Ciência Política	1	-0.1317	0.1818	0.5245	0.4689
curso2	Ciências Contábeis	1	-0.3261	0.1563	4.3504	0.0370
curso2	Ciências Sociais	1	0.4656	0.1872	6.1862	0.0129
curso2	Geografia	1	0.2134	0.1729	1.5245	0.2169
curso2	Outros	1	-0.7091	0.2368	8.9660	0.0028
curso2	Relações Internacion	1	-0.5579	0.1731	10.3895	0.0013
dataingressocateg	2001	1	0.4265	0.2206	3.7363	0.0532
dataingressocateg	2002	1	0.2470	0.1899	1.6913	0.1934
dataingressocateg	2003	1	0.3824	0.1431	7.1425	0.0075
dataingressocateg	2004	1	-0.0129	0.1319	0.0096	0.9221
dataingressocateg	2005	1	-0.0706	0.1199	0.3470	0.5558
dataingressocateg	2006	1	-0.3237	0.1393	5.4007	0.0201
dataingressocateg	2007	1	-0.6833	0.1831	13.9295	0.0002
dataingressocateg	2008	1	-0.7768	0.2577	9.0869	0.0026
Nome_Professor	2	1	1.9712	28.8268	0.0047	0.9455
Nome_Professor	3	1	-0.00736	28.8338	0.0000	0.9998
Nome_Professor	6	1	1.6216	28.8255	0.0032	0.9551
Nome_Professor	8	1	2.1414	28.8261	0.0055	0.9408
Nome_Professor	10	1	0.4294	28.8288	0.0002	0.9881
Nome_Professor	11	1	-0.0565	28.8276	0.0000	0.9984
Nome_Professor	14	1	-12.5199	588.5	0.0005	0.9830
Nome_Professor	18	1	-13.8502	364.8	0.0014	0.9697
Nome_Professor	22	1	2.7298	28.8261	0.0090	0.9246
Nome_Professor	23	1	1.1030	28.8257	0.0015	0.9695
Nome_Professor	25	1	0.7573	28.8253	0.0007	0.9790
Nome_Professor	27	1	-0.0299	28.8266	0.0000	0.9992
Nome_Professor	28	1	1.5703	28.8272	0.0030	0.9566
Nome_Professor	29	1	2.0740	28.8254	0.0052	0.9426
Nome_Professor	30	1	0.4400	28.8266	0.0002	0.9878
Nome_Professor	31	1	1.4986	28.8269	0.0027	0.9585
Nome_Professor	33	1	-0.0789	28.8270	0.0000	0.9978
Nome_Professor	34	1	2.9989	28.8257	0.0108	0.9171
Nome_Professor	36	1	2.9088	28.8260	0.0102	0.9196
Nome_Professor	37	1	0.6873	28.8257	0.0006	0.9810
Nome_Professor	38	1	1.5779	28.8296	0.0030	0.9564
Nome_Professor	39	1	1.5732	28.8256	0.0030	0.9565
Nome_Professor	40	1	0.7361	28.8267	0.0007	0.9796
Nome_Professor	41	1	0.7176	28.8288	0.0006	0.9801
Turno	noturno	1	0.5170	0.2105	6.0330	0.0140
Turno	vespertino	1	0.0330	0.1719	0.0369	0.8477
professor_quadro	não	1	-0.3378	0.1436	5.5340	0.0187
PERIODO	20042	1	-0.4527	0.2435	3.4562	0.0630
PERIODO	20051	1	-0.9122	0.1987	21.0745	<.0001
PERIODO	20052	1	-0.3891	0.1659	5.5029	0.0190

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
PERIODO	20061	1	-0.1804	0.1621	1.2376	0.2659
PERIODO	20062	1	-0.2824	0.1702	2.7524	0.0971
PERIODO	20071	1	0.2438	0.1620	2.2660	0.1322
PERIODO	20072	1	0.3059	0.2108	2.1063	0.1467
PERIODO	20081	1	0.9590	0.2058	21.7141	<.0001
PERIODO	20082	1	1.1487	0.2579	19.8436	<.0001

Apêndice 7 – Parâmetros estimados para o modelo da disciplina **BIO.** tendo como variável resposta a reprovação presencial:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2391	0.1954	131.3054	<.0001
SEXO	F	1	-0.3561	0.0955	13.8971	0.0002
fez_no_fluxo	nao	1	0.5404	0.1993	7.3563	0.0067
fez_no_fluxo	outro	1	0.1552	0.3603	0.1855	0.6667
curso2	Agronomia	1	0.0500	0.1941	0.0662	0.7969
curso2	Engenharia Florestal	1	0.7875	0.1948	16.3501	<.0001
curso2	Medicina Veterinária	1	0.1673	0.2115	0.6261	0.4288
dataingressocateg	2001	1	0.9517	0.4434	4.6076	0.0318
dataingressocateg	2002	1	0.6915	0.3259	4.5017	0.0339
dataingressocateg	2003	1	0.6391	0.2278	7.8738	0.0050
dataingressocateg	2004	1	0.3843	0.2453	2.4542	0.1172
dataingressocateg	2005	1	-0.2588	0.2648	0.9553	0.3284
dataingressocateg	2006	1	-0.8357	0.3065	7.4353	0.0064
dataingressocateg	2007	1	-1.1291	0.4113	7.5356	0.0060
dataingressocateg	2008	1	-1.0904	0.6111	3.1842	0.0744
Nome_Professor	5	1	0.5847	0.5486	1.1357	0.2866
Nome_Professor	11	1	0.1878	0.4350	0.1865	0.6659
Nome_Professor	14	1	1.6674	0.4172	15.9706	<.0001
Nome_Professor	15	1	-0.00907	0.2818	0.0010	0.9743
Nome_Professor	17	1	0.8802	0.2292	14.7463	0.0001
Nome_Professor	20	1	-3.8710	0.9664	16.0440	<.0001
Nome_Professor	22	1	-0.4414	0.5605	0.6201	0.4310
Nome_Professor	29	1	-0.4436	0.3364	1.7380	0.1874
Nome_Professor	32	1	-0.5343	0.4769	1.2548	0.2626
PERIODO	20042	1	0.3612	0.2952	1.4969	0.2211
PERIODO	20051	1	-1.7638	0.3935	20.0913	<.0001
PERIODO	20052	1	-1.3615	0.3091	19.4071	<.0001
PERIODO	20061	1	-0.6163	0.3193	3.7268	0.0535
PERIODO	20062	1	-0.1920	0.3085	0.3874	0.5337
PERIODO	20071	1	0.1432	0.3381	0.1793	0.6720
PERIODO	20072	1	-0.1653	0.3568	0.2147	0.6431
PERIODO	20081	1	1.2096	0.3681	10.7984	0.0010
PERIODO	20082	1	1.6845	0.5341	9.9477	0.0016

Apêndice 8 – Parâmetros estimados para o modelo da disciplina **PE.** tendo como variável resposta a reprovação presencial:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0122	0.3214	87.8570	<.0001
SEXO	F	1	-0.2226	0.0760	8.5734	0.0034
fez_no_fluxo	nao	1	0.3620	0.2883	1.5768	0.2092
fez_no_fluxo	outro	1	0.0809	0.5685	0.0203	0.8868
n_SR		1	0.5157	0.1645	9.8312	0.0017
ordem_q_cursou	2vez	1	0.3667	0.2025	3.2812	0.0701
ordem_q_cursou	3vez	1	-0.2069	0.2506	0.6811	0.4092
ordem_q_cursou	4+vez	1	-1.2466	0.4973	6.2825	0.0122
curso2	Ciência da Computaçã	1	0.2162	0.1693	1.6307	0.2016
curso2	Computação	1	0.3965	0.1751	5.1305	0.0235
curso2	Engenharia Civil	1	0.3078	0.1519	4.1078	0.0427
curso2	Engenharia Elétrica	1	-0.4201	0.1970	4.5496	0.0329
curso2	Engenharia Mecatrôni	1	-0.3038	0.2056	2.1827	0.1396
curso2	Engenharia Mecânica	1	-0.7753	0.1976	15.3935	<.0001
curso2	Engenharia de Redes	1	0.0704	0.1861	0.1429	0.7054
curso2	Matemática	1	0.7044	0.1649	18.2534	<.0001
dataingressocateg	2001	1	0.4427	0.2931	2.2812	0.1310
dataingressocateg	2002	1	0.1874	0.2193	0.7304	0.3928
dataingressocateg	2003	1	0.1424	0.1808	0.6202	0.4310
dataingressocateg	2004	1	0.1128	0.1460	0.5970	0.4397
dataingressocateg	2005	1	-0.2449	0.1417	2.9896	0.0838
dataingressocateg	2006	1	-0.2079	0.1518	1.8751	0.1709
dataingressocateg	2007	1	-0.2524	0.1852	1.8566	0.1730
dataingressocateg	2008	1	-1.0328	0.2531	16.6557	<.0001
Nome_Professor	1	1	0.1926	0.2345	0.6746	0.4115
Nome_Professor	2	1	-0.5467	0.4247	1.6567	0.1980
Nome_Professor	5	1	0.6016	0.3503	2.9502	0.0859
Nome_Professor	7	1	0.8643	0.1982	19.0215	<.0001
Nome_Professor	13	1	0.8477	0.2625	10.4268	0.0012
Nome_Professor	14	1	-0.4121	0.3132	1.7320	0.1882
Nome_Professor	17	1	-0.6801	0.2665	6.5138	0.0107
Nome_Professor	18	1	-0.1672	1.0228	0.0267	0.8702
Nome_Professor	25	1	0.0777	0.2103	0.1366	0.7117
Nome_Professor	26	1	-0.7467	0.3643	4.2004	0.0404
Nome_Professor	27	1	-0.1274	0.2956	0.1856	0.6666
Nome_Professor	28	1	1.1759	0.3378	12.1142	0.0005
Nome_Professor	30	1	-1.6800	0.4334	15.0282	0.0001
Nome_Professor	35	1	2.0174	0.2244	80.8201	<.0001
Nome_Professor	41	1	-0.8914	0.3659	5.9340	0.0149
PERIODO	20042	1	-0.9429	0.2543	13.7485	0.0002
PERIODO	20051	1	-1.0397	0.2359	19.4333	<.0001
PERIODO	20052	1	-1.0677	0.2499	18.2561	<.0001
PERIODO	20061	1	-0.1624	0.2309	0.4951	0.4817
PERIODO	20062	1	0.1110	0.1795	0.3822	0.5364
PERIODO	20071	1	1.4402	0.2454	34.4382	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
PERIODO	20072	1	0.6603	0.2285	8.3533	0.0038
PERIODO	20081	1	1.0846	0.2415	20.1676	<.0001
PERIODO	20082	1	1.7695	0.2896	37.3312	<.0001

Apêndice 9 – *Odds ratio* do modelo de reprovação Geral para o modelo Completo

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SEXO F vs M	0.592	0.530	0.661
DISCIPLINA BIO vs PE	0.392	0.258	0.593
DISCIPLINA EA vs PE	0.554	0.377	0.814
brasileiro nao vs sim	2.499	1.883	3.315
fez_no_fluxo nao vs sim	1.987	1.758	2.246
fez_no_fluxo outro vs sim	0.835	0.146	4.776
MODALIDADE ML vs OBR	4.658	2.928	7.409
MODALIDADE OPT vs OBR	0.876	0.665	1.154
ordem_q_cursou 2vez vs 1vez	1.037	0.894	1.204
ordem_q_cursou 3vez vs 1vez	0.585	0.448	0.764
ordem_q_cursou 4+vez vs 1vez	0.534	0.331	0.862
curso2 Administração Diurno vs Sociologia	0.719	0.499	1.035
curso2 Administração Noturn vs Sociologia	0.829	0.567	1.213
curso2 Agronomia vs Sociologia	0.964	0.583	1.595
curso2 Antropologia vs Sociologia	1.158	0.756	1.773
curso2 Arquivologia vs Sociologia	1.074	0.741	1.555
curso2 Biblioteconomia vs Sociologia	2.114	1.454	3.073
curso2 Ciência Política vs Sociologia	0.763	0.513	1.137
curso2 Ciência da Computaçã vs Sociologia	0.474	0.285	0.790
curso2 Ciências Contábeis vs Sociologia	0.758	0.529	1.084
curso2 Ciências Sociais vs Sociologia	1.501	1.009	2.232
curso2 Computação vs Sociologia	0.698	0.418	1.166
curso2 Engenharia Civil vs Sociologia	0.542	0.331	0.886
curso2 Engenharia Elétrica vs Sociologia	0.292	0.170	0.501
curso2 Engenharia Florestal vs Sociologia	1.096	0.662	1.814
curso2 Engenharia Mecatrôni vs Sociologia	0.330	0.188	0.580
curso2 Engenharia Mecânica vs Sociologia	0.249	0.148	0.418
curso2 Engenharia de Redes vs Sociologia	0.497	0.291	0.848
curso2 Geografia vs Sociologia	1.353	0.921	1.988
curso2 Matemática vs Sociologia	1.059	0.645	1.740
curso2 Medicina Veterinária vs Sociologia	0.705	0.408	1.219
curso2 Outros vs Sociologia	1.087	0.183	6.459
curso2 Relações Internacion vs Sociologia	0.432	0.292	0.637
dataingressocateg 2001 vs anterior a 2001	0.903	0.668	1.220
dataingressocateg 2002 vs anterior a 2001	0.741	0.564	0.974
dataingressocateg 2003 vs anterior a 2001	0.719	0.566	0.914
dataingressocateg 2004 vs anterior a 2001	0.506	0.395	0.649

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
dataingressocateg 2005 vs anterior a 2001	0.413	0.317	0.538
dataingressocateg 2006 vs anterior a 2001	0.345	0.258	0.462
dataingressocateg 2007 vs anterior a 2001	0.284	0.204	0.394
dataingressocateg 2008 vs anterior a 2001	0.200	0.136	0.296
Nome_Professor 1 vs 43	1.515	0.763	3.008
Nome_Professor 2 vs 43	1.502	0.739	3.053
Nome_Professor 3 vs 43	0.166	0.050	0.554
Nome_Professor 5 vs 43	0.880	0.420	1.842
Nome_Professor 6 vs 43	1.511	0.761	3.002
Nome_Professor 7 vs 43	2.424	1.294	4.541
Nome_Professor 8 vs 43	1.387	0.668	2.880
Nome_Professor 10 vs 43	1.238	0.523	2.931
Nome_Professor 11 vs 43	0.693	0.343	1.399
Nome_Professor 13 vs 43	0.746	0.383	1.453
Nome_Professor 14 vs 43	0.976	0.493	1.932
Nome_Professor 15 vs 43	1.429	0.683	2.987
Nome_Professor 17 vs 43	1.564	0.833	2.938
Nome_Professor 18 vs 43	0.232	0.095	0.567
Nome_Professor 20 vs 43	0.492	0.205	1.184
Nome_Professor 22 vs 43	3.122	1.487	6.551
Nome_Professor 23 vs 43	0.634	0.325	1.238
Nome_Professor 25 vs 43	1.210	0.640	2.289
Nome_Professor 26 vs 43	0.861	0.383	1.936
Nome_Professor 27 vs 43	1.038	0.522	2.063
Nome_Professor 28 vs 43	0.587	0.294	1.172
Nome_Professor 29 vs 43	1.047	0.561	1.955
Nome_Professor 30 vs 43	0.203	0.103	0.402
Nome_Professor 31 vs 43	0.547	0.252	1.186
Nome_Professor 32 vs 43	0.522	0.206	1.323
Nome_Professor 33 vs 43	0.687	0.341	1.385
Nome_Professor 34 vs 43	1.658	0.866	3.176
Nome_Professor 35 vs 43	5.403	2.822	10.345
Nome_Professor 36 vs 43	1.618	0.807	3.245
Nome_Professor 37 vs 43	0.622	0.315	1.229
Nome_Professor 38 vs 43	0.881	0.341	2.276
Nome_Professor 39 vs 43	0.559	0.292	1.070
Nome_Professor 40 vs 43	0.822	0.418	1.618
Nome_Professor 41 vs 43	0.267	0.127	0.560
Nome_Professor 42 vs 43	0.105	0.038	0.293
Turno noturno vs matutino	1.100	0.865	1.399
Turno vespertino vs matutino	0.682	0.544	0.854
PERIODO 20042 vs 20041	1.159	0.906	1.482
PERIODO 20051 vs 20041	0.797	0.611	1.040
PERIODO 20052 vs 20041	0.862	0.660	1.127
PERIODO 20061 vs 20041	1.457	1.110	1.914
PERIODO 20062 vs 20041	1.973	1.487	2.618
PERIODO 20071 vs 20041	2.342	1.731	3.170

Odds Ratio Estimates				
Effect		Point Estimate	95% Wald Confidence Limits	
PERIODO	20072 vs 20041	1.585	1.134	2.216
PERIODO	20081 vs 20041	2.855	2.032	4.011
PERIODO	20082 vs 20041	3.250	2.232	4.732

Apêndice 10 – *Odds ratio* do modelo de reprovação presencial da disciplina **E.A.**

Odds Ratio Estimates				
Effect		Point Estimate	95% Wald Confidence Limits	
SEXO	F vs M	0.657	0.547	0.788
brasileiro	nao vs sim	4.565	3.161	6.593
MODALIDADE	ML vs OBR	6.384	2.489	16.373
MODALIDADE	OPT vs OBR	0.724	0.313	1.677
n_SR		1.516	1.138	2.018
ordem_q_cursou	2vez vs 1vez	1.251	0.944	1.658
ordem_q_cursou	3vez vs 1vez	0.323	0.177	0.590
ordem_q_cursou	4+vez vs 1vez	0.395	0.148	1.058
curso2	Administração Diurno vs Sociologia	0.667	0.423	1.049
curso2	Administração Noturn vs Sociologia	0.504	0.294	0.864
curso2	Agronomia vs Sociologia	0.605	0.109	3.371
curso2	Antropologia vs Sociologia	0.973	0.561	1.688
curso2	Arquivologia vs Sociologia	1.171	0.720	1.905
curso2	Biblioteconomia vs Sociologia	1.807	1.119	2.919
curso2	Ciência Política vs Sociologia	0.667	0.403	1.104
curso2	Ciências Contábeis vs Sociologia	0.549	0.342	0.882
curso2	Ciências Sociais vs Sociologia	1.212	0.724	2.027
curso2	Geografia vs Sociologia	0.942	0.577	1.537
curso2	Outros vs Sociologia	0.374	0.201	0.696
curso2	Relações Internacion vs Sociologia	0.435	0.268	0.707
dataingressocateg	2001 vs anterior a 2001	0.681	0.395	1.174
dataingressocateg	2002 vs anterior a 2001	0.569	0.341	0.948
dataingressocateg	2003 vs anterior a 2001	0.651	0.431	0.983
dataingressocateg	2004 vs anterior a 2001	0.439	0.286	0.672
dataingressocateg	2005 vs anterior a 2001	0.414	0.267	0.643
dataingressocateg	2006 vs anterior a 2001	0.321	0.195	0.531
dataingressocateg	2007 vs anterior a 2001	0.224	0.126	0.400
dataingressocateg	2008 vs anterior a 2001	0.204	0.101	0.415
Nome_Professor	2 vs 42	19.393	3.368	111.661
Nome_Professor	3 vs 42	2.681	0.310	23.215
Nome_Professor	6 vs 42	13.671	2.595	72.014
Nome_Professor	8 vs 42	22.990	4.347	121.582
Nome_Professor	10 vs 42	4.150	0.668	25.786
Nome_Professor	11 vs 42	2.553	0.386	16.869
Nome_Professor	14 vs 42	<0.001	<0.001	>999.999
Nome_Professor	18 vs 42	<0.001	<0.001	>999.999
Nome_Professor	22 vs 42	41.410	7.419	231.140

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Nome_Professor 23 vs 42	8.139	1.585	41.782
Nome_Professor 25 vs 42	5.760	1.035	32.068
Nome_Professor 27 vs 42	2.622	0.444	15.482
Nome_Professor 28 vs 42	12.987	2.203	76.564
Nome_Professor 29 vs 42	21.492	4.336	106.524
Nome_Professor 30 vs 42	4.194	0.800	21.979
Nome_Professor 31 vs 42	12.089	2.225	65.698
Nome_Professor 33 vs 42	2.496	0.433	14.396
Nome_Professor 34 vs 42	54.197	11.669	251.724
Nome_Professor 36 vs 42	49.526	10.311	237.886
Nome_Professor 37 vs 42	5.371	1.003	28.765
Nome_Professor 38 vs 42	13.086	1.965	87.138
Nome_Professor 39 vs 42	13.025	2.944	57.618
Nome_Professor 40 vs 42	5.639	1.070	29.726
Nome_Professor 41 vs 42	5.536	0.990	30.962
Turno noturno vs matutino	2.907	1.365	6.191
Turno vespertino vs matutino	1.792	0.951	3.376
professor_quadro não vs sim	0.509	0.290	0.893
PERIODO 20042 vs 20041	0.988	0.582	1.677
PERIODO 20051 vs 20041	0.624	0.372	1.047
PERIODO 20052 vs 20041	1.053	0.642	1.726
PERIODO 20061 vs 20041	1.297	0.782	2.152
PERIODO 20062 vs 20041	1.172	0.673	2.039
PERIODO 20071 vs 20041	1.983	1.141	3.444
PERIODO 20072 vs 20041	2.110	1.091	4.080
PERIODO 20081 vs 20041	4.054	2.101	7.823
PERIODO 20082 vs 20041	4.901	2.312	10.389

Apêndice 11 – *Odds ratio* do modelo de reprovação presencial da disciplina **P.E.**

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SEXO F vs M	0.641	0.476	0.863
fez_no_fluxo nao vs sim	2.236	1.737	2.879
fez_no_fluxo outro vs sim	1.688	0.314	9.082
n_SR	1.675	1.213	2.312
ordem_q_cursou 2vez vs 1vez	0.487	0.346	0.685
ordem_q_cursou 3vez vs 1vez	0.274	0.147	0.512
ordem_q_cursou 4+vez vs 1vez	0.097	0.025	0.378
curso2 Ciência da Computaçã vs Outros	1.510	0.290	7.859
curso2 Computação vs Outros	1.809	0.345	9.472
curso2 Engenharia Civil vs Outros	1.655	0.321	8.537
curso2 Engenharia Elétrica vs Outros	0.799	0.151	4.223
curso2 Engenharia Mecatrôni vs Outros	0.898	0.169	4.778
curso2 Engenharia Mecânica vs Outros	0.560	0.105	2.989

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
curso2 Engenharia de Redes vs Outros	1.305	0.248	6.857
curso2 Matemática vs Outros	2.461	0.474	12.785
dataingressocateg 2001 vs anterior a 2001	0.664	0.291	1.514
dataingressocateg 2002 vs anterior a 2001	0.514	0.251	1.054
dataingressocateg 2003 vs anterior a 2001	0.492	0.250	0.966
dataingressocateg 2004 vs anterior a 2001	0.477	0.242	0.941
dataingressocateg 2005 vs anterior a 2001	0.334	0.166	0.670
dataingressocateg 2006 vs anterior a 2001	0.346	0.167	0.720
dataingressocateg 2007 vs anterior a 2001	0.331	0.150	0.729
dataingressocateg 2008 vs anterior a 2001	0.152	0.063	0.366
Nome_Professor 1 vs 43	2.051	0.784	5.364
Nome_Professor 2 vs 43	0.979	0.288	3.332
Nome_Professor 5 vs 43	3.087	0.994	9.591
Nome_Professor 7 vs 43	4.014	1.641	9.819
Nome_Professor 13 vs 43	3.948	1.435	10.863
Nome_Professor 14 vs 43	1.120	0.396	3.171
Nome_Professor 17 vs 43	0.857	0.384	1.914
Nome_Professor 18 vs 43	1.431	0.143	14.356
Nome_Professor 25 vs 43	1.828	0.788	4.240
Nome_Professor 26 vs 43	0.802	0.260	2.468
Nome_Professor 27 vs 43	1.489	0.520	4.263
Nome_Professor 28 vs 43	5.482	1.826	16.456
Nome_Professor 30 vs 43	0.315	0.111	0.895
Nome_Professor 35 vs 43	12.716	4.781	33.826
Nome_Professor 41 vs 43	0.694	0.233	2.063
PERIODO 20042 vs 20041	2.484	1.365	4.520
PERIODO 20051 vs 20041	2.255	1.085	4.684
PERIODO 20052 vs 20041	2.193	0.955	5.033
PERIODO 20061 vs 20041	5.421	2.340	12.561
PERIODO 20062 vs 20041	7.126	3.119	16.282
PERIODO 20071 vs 20041	26.923	10.415	69.598
PERIODO 20072 vs 20041	12.343	4.877	31.240
PERIODO 20081 vs 20041	18.866	7.324	48.595
PERIODO 20082 vs 20041	37.423	14.055	99.637

Anexos

Anexo 1A – Ementa da disciplina Probabilidade e Estatística.

- Análise de observações;
- Modelo matemático;
- Exp. Aleatória e espaço amostral, axiomas e teoremas básicos;
- Variáveis aleatórias;
- Distribuições e suas características;
- Covariância e Correlação;
- Distribuição conjunta;
- Principais modelos – discretos e contínuos;
- Estatística descritiva;
- Ajustamentos de funções reais;
- Correlação e regressão;
- Noções de amostragem;
- Testes de hipóteses e;
- Aplicações.

Anexo 1B – Programa da disciplina Probabilidade e Estatística.

<u>Unidade I – Fundamentos do cálculo de probabilidade</u>

- Conceitos e definições;
- Axiomas e teoremas básicos;
- Probabilidade condicionada e eventos independentes;
- Experiência aleatória uniforme.

<u>Unidade II – Variáveis aleatórias e suas distribuições</u>

- Variável aleatória: definição;
- Variável aleatória unidimensional;
- Variável aleatória bidimensional.

<u>Unidade III – Medidas características de uma distribuição de probabilidade</u>

- Expectância e suas propriedades;
- Momentos e suas funções;
- Separatrizes;
- Moda.

<u>Unidade IV – Modelos probabilísticos</u>

- Distribuições unidimensionais do tipo discreto: Bernoulli, Binomial, Poisson, Geométrica e Hipergeométrica;
- Distribuições unidimensionais do tipo contínuo: Uniforme, Normal, Exponencial, Qui-quadrado, Student.

Unidade V – Análise estatística

- Distribuição de frequência;
- Medidas características das distribuições: posição, dispersão, assimetria e curtose;
- Ajustamento de um modelo probabilístico a uma distribuição de frequência;
- Correlação e regressão linear.

Unidade VI – Análise Dinâmica de observações

- Séries temporais;
- Ajustamento de uma função real a uma série temporal.

Unidade VII – Noções de amostragem e estimação

- População e população matriz, censo e amostragem;
- Amostra aleatória, estimador e estimativa;
- Intervalos de confiança para a média, o total e a proporção.

Unidade VIII – Noções de testes de hipóteses

- Formulação geral de um teste paramétrico;
- Estudo de alguns testes paramétricos: médias e proporções;
- Teste Qui-quadrado.

Anexo 2A – Ementa da disciplina Estatística Aplicada.

- Conceitos básicos;
- Distribuição de frequências e suas características;
- Introdução a probabilidade;
- Ajustamento de funções reais;
- Correlação e regressão linear;
- Noções de amostragem e;
- Testes de hipóteses.

Anexo 2B – Programa da disciplina Estatística Aplicada.

Unidade I – Conceitos básicos

- Estatística indutiva e dedutiva;
- Variáveis e constantes, cálculo de somatórios;
- Frequências, porcentagens e proporções;
- Realidade e modelo.

Unidade II – Distribuições de frequências

- Rol e distribuição de frequências;
- Tipos de frequências: simples, relativas e acumuladas;
- Histograma e polígono de frequências – análise gráfica;

Unidade III – Medidas de tendência central e separatrizes

- Conceituação, média aritmética, moda e mediana;

- Separatrizes, aplicações.

Unidade IV – Medidas de dispersão, assimetria e curtose

- Variância, desvio padrão e coeficiente de variação;
- Momentos ordinários e centrais, coeficientes.

Unidade V – Introdução a probabilidade

- Conjunto, espaço-amostra a eventos, eventos mutuamente exclusivos e independentes, eventos complementares, união e interpretação de eventos;
- Variável aleatória, distribuição de probabilidade;
- Modelos discretos e contínuos, uso de tabelas;
- Aplicações em ajustamentos de modelos probabilísticos.

Unidade VI – Ajustamento de funções reais e séries temporais

- Séries temporais, tendência de uma série temporal;
- Método dos mínimos quadrados, ajustamento de polinômios e exponencial, saturações, aplicações.

Unidade VII – Correlação e regressão

- Conceitos básicos, coeficiente de correlação linear;
- Retas de regressão, proporção de variação explicada, outros coeficientes de regressão, aplicações.

Unidade VIII – Noções de amostragem e testes de hipóteses

- Amostras aleatórias, distribuições de amostragem;
- Estimação de características, dimensionamento da amostra;
- Formulação geral de um teste;
- Principais testes paramétricos e não paramétricos;
- Elaboração de trabalho final.

Anexo 3A – Ementa da disciplina Bioestatística**.**

- Estatística descritiva;
- Noções de probabilidade;
- Principais modelos discretos e contínuos;
- Ajustamento de modelos probabilísticos;
- Noções de amostragem e estimação;
- Noções de testes de hipóteses;
- Análise de variância: classificação simples;
- Correlação e regressão linear e;
- Noções sobre experimentos e levantamentos.

Anexo 3B – Programa da disciplina Bioestatística.

Unidade I – Estatística descritiva

- Escalas e medidas;

- Apresentação de dados numéricos: tabelas e gráficos;
- Distribuição de frequências - histograma e polígono de frequências;
- Medidas de posição: média aritmética, mediana, moda, separatrizes e outras medidas;
- Medidas de dispersão: variância, desvio padrão e coeficiente de variação;
- Assimetria e curtose;
- Coeficiente de correlação linear.

Unidade II – Noções de probabilidade

- Experiência aleatória, espaço-amostra, eventos e axiomas e teoremas básicos;
- Variáveis aleatórias discretas e contínuas: expectância e momentos;
- Principais modelos probabilísticos: Binomial, Poisson, Normal, Qui-quadrado, Student e F. Uso de tabela e aplicações;

Unidade III – Noções de amostragem e estimação

- População e amostra, censo e amostragem;
- Amostra aleatória, estimador e estimativa;
- Dimensionamento da amostra;
- Intervalo de confiança

Unidade IV – Noções de teste de hipóteses

- Formulação geral de um teste paramétrico, tipos de erro;
- Testes de médias e variâncias;
- Teste qui-quadrado
- Análise de variância: classificação simples, testes de comparações múltiplas, repetição e casualização.

Unidade V – Correlação e regressão linear

- Correlação e regressão linear simples;
- Análise de variância na regressão.

Referências

- AGRESTI, ALAN. **An introduction to Categorical Data Analysis**, John Wiley & Sons, Inc., New York, 1996.
- HOSMER, DAVID W.; LEMESHOW, STANLEY. **Applied logistic regression**. 2a edição. New York: John Wiley.
- NETER, J., KUTNER, M., NACHTSHEIM, C. J. e WASSERMAN, W. **Applied linear statistical models**. 5a edição. Illinois: Irwin, 2005.
- OLIVEIRA, J. S. R. **Modelos de Regressão Logística**.2011.84f. Monografia (Bacharelado em Estatística) – Universidade de Brasília, Brasília.
- PEREIRA, Ângela R.de A., GOMES, Marília M. F., et al. **Mortalidade de crianças e adolescentes no Distrito Federal e região do Entorno(RIDE): condições de vida e desigualdades intra-urbanas**. In: Encontro Nacional de Estudos Populacionais,14,2004, Caxambú-MG
- VASCONCELOS, Ana Maria.N., FERREIRA, Ignez C. B., et al. **Da utopia à realidade: uma análise dos fluxos migratórios para o Aglomerado Urbano de Brasília**. In: Encontro Nacional de Estudos Populacionais,15,2006, Caxambú-MG